Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte



Impressum

Herausgeber:	BITKOM Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V. Albrechtstraße 10 A, 10117 Berlin-Mitte Tel.: 030.27576-0 Fax: 030.27576-400 bitkom@bitkom.org www.bitkom.org
Ansprechpartner:	Dr. Mathias Weber Tel.: 030.27576-121 m.weber@bitkom.org
Verantwortliches Gremium:	BITKOM-Arbeitskreis Big Data
Projektleitung:	Jürgen Urbanski (T-Systems International GmbH) Dr. Mathias Weber (BITKOM)
Copyright:	BITKOM 2012
Grafik/Layout:	Design Bureau kokliko/ Astrid Scheibe (BITKOM)
Titelbild:	© Ben Chams – Fotolia.com

Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte



Inhaltsverzeichnis

Verzeichnis der Tabellen 4
Verzeichnis der Abbildungen 4
1 Geleitwort 5
2 Management Summary 7
3 Big Data – Chancen und Herausforderungen für Unternehmen 11
3.1 Mit Big Data verbundene Chancen 14
3.2 Mit Big Data verbundene Herausforderungen 15
4 Big Data – Begriffsbestimmung 19
5 Einordnung von Big Data in die Entwicklungslinien der Technologien und Transformationsstrategien 22
5.1 Transaktionale Systeme 23
5.2 Analytische Systeme: Data Warehouse und Business Intelligence 24
5.3 Dokumentenmanagement 24
5.4 Auswirkungen auf die Software-Entwicklung 25
5.5 Auswirkungen auf die Anwendungs-Architektur 27
5.6 Big Data als Fortentwicklung vorhandener Technologien 28
5.7 Auswirkungen von Big Data auf benachbarte Disziplinen 29
6 Nutzung unstrukturierter Daten in Big-Data-Verfahren 30
7 Big Data – Praxiseinsatz und wirtschaftlicher Nutzen 34
7.1 Marketing & Vertrieb 35
7.2 Forschung und Produktentwicklung 36
7.3 Produktion, Service und Support 38
7.4 Distribution und Logistik 39
7.5 Finanz- und Risiko-Controlling 41
8 Big Data und Datenschutz 43
9 Big Data – Marktentwicklung in wichtigen Regionen 47
10 Einsatzbeispiele von Big Data in Wirtschaft und Verwaltung 51
10.1 Einsatzbeispiele aus Marketing und Vertrieb 54
10.1.1 (№01) Deutsche Welle – Nutzungsdaten und -analysen von Web-Videos auf einen Blick 54
10.1.2 (№02) DeutschlandCard GmbH – Effiziente IT-Infrastruktur für Big Data 56
10.1.3 (№03) dm – Mitarbeitereinsatzplanung 58
10.1.4 (№04) etracker – Verbindung konventioneller IT-Systeme mit Big-Data-Technologien 59
10.1.5 (№05) Macy's – Preisoptimierung 60
10.1.6 (№06) MZinga – Echtzeit-Analysen von »Social Intelligence« 61
10.1.7 (№07) Otto – Verbesserung der Absatzprognose 62
10.1.8 (№08) Satelliten-TV Anbieter – Customer Churn und »Pay-per-View«-Werbeoptimierung (Pilot) 63
10.1.9 (№09) Searchmetrics – Realtime-Abfragen und Auswertungen auf Milliarden von Datensätzen 64
10.1.10 (№10) Schukat Electronic – Live-Analyse von Auftragsdurchlaufzeiten im Dashboard 65
10.1.11 (№11) Telecom Italia – Minimierung der Kundenfluktuation 66
10.1.12 (№12) Webtrekk GmbH – Realtime-Webanalyse 67

10.2 Einsatzbeispiele aus Forschung und Entwicklung 68
10.2.1 (N°13) Mittelständische Unternehmensberatung – Competitive Intelligence: Trend-Analyse im Internet 68
10.2.2 (N°14) Königliches Technologie Institut Stockholm – Realzeit-Analyse für Smarter Cities 69
10.2.3 (N°15) University of Ontario – Verarbeitung von Sensordaten medizinischer Überwachungsgeräte 70
10.3 Einsatzbeispiele aus der Produktion 71
10.3.1 (N°16) Energietechnik – Überwachung und Diagnose bei komplexen technischen Systemen 71
10.3.2 (N°17) Semikron GmbH – Geschäftsprozesse optimieren durch ganzheitliches Mess- und Prozessdatenmanagement 72
10.3.3 (N°18) Vaillant – Globale Planung und Steuerung bis auf Produktebene 74
10.4 Einsatzbeispiele aus Service und Support 75
10.4.1 (N°19) Automobilhersteller – Ganzheitliche Qualitätsanalyse durch integrierte Daten 75
10.4.2 (N°20) Treato – Analyse von unstrukturierten Gesundheitsdaten 76
10.5 Einsatzbeispiel aus Distribution und Logistik 77
10.5.1 (N°21) TomTom Business Solutions – Flottenmanagement in Echtzeit 77
10.6 Einsatzbeispiele aus Finanz- und Risiko-Controlling 78
10.6.1 (N°22) Europäischer Spezialist für Kreditkartensicherheit – Kreditkartenbetrugsanalyse 78
10.6.2 (N°23) Paymint AG – Fraud Detection in Kreditkartentransaktionen 79
10.6.3 (N°24) United Overseas Bank (Singapur) – Risikoabschätzung in Echtzeit 80
10.7 Einsatzbeispiele aus Administration, Organisation und Operations 81
10.7.1 (N°25) Aadhaar-Projekt – Personenidentifikation indischer Bürger als Grundlage für Verwaltungs- und Geschäftsprozesse 81
10.7.2 (N°26) Anbieter für Dialog-Marketing per E-Mail – Plattform für Versand vertrauenswürdiger E-Mails 82
10.7.3 (N°27) Paketdienstleister – Sicherung der Compliance 84
10.7.4 (N°28) Expedia – Prozessoptimierung bringt 25fachen ROI 85
10.7.5 (N°29) NetApp – Diagnoseplattform 86
10.7.6 (N°30) Otto-Gruppe – Mehr Sicherheit und Qualität für die gesamte IT 88
10.7.7 (N°31) Europäisches Patentamt – Patentrecherche weltweit 89
10.7.8 (N°32) Schweizer Staatssekretariat für Wirtschaft – Kostengünstige Höchstleistung für die Schweizer Arbeitsmarktstatistik 90
10.7.9 (N°33) Toll Collect – Qualitätssicherung der automatischen Mauterhebung 91
10.7.10 (N°34) XING AG – Bewältigung schnell wachsender Datenvolumina 92
11 Abkürzungen und Glossar 93
12 Sachwortregister 97
Autoren des Leitfadens 101



Verzeichnis der Tabellen

Tabelle 1: Chancen durch Big Data 14
Tabelle 2: Mit Big Data verbundene Herausforderungen an Unternehmen 17
Tabelle 3: Facetten von Big Data 21
Tabelle 4: Vergleich der Schwerpunkte transaktionaler Systeme und Big Data 23
Tabelle 5: Analytische Systeme und Big Data – Vergleich der Schwerpunkte 24
Tabelle 6: Dokumenten-Management-Systeme und Big Data – Vergleich der Schwerpunkte 25
Tabelle 7: Schwerpunkte bei der Anwendungsanalyse 27
Tabelle 8: Schritte in Richtung Big Data 29
Tabelle 9: Transformationspotenzial durch Big Data 50
Tabelle 10: Funktionsbereiche der Big-Data-Einsatzbeispiele 51
Tabelle 11: Wirtschaftszweige der Big-Data-Einsatzbeispiele 51
Tabelle 12: Einsatzbeispiele für Big-Data – Übersicht nach Funktionsbereichen und Wirtschaftszweigen 52

Verzeichnis der Abbildungen

Abbildung 1: Welche Informationstechnologien das Big-Data-Phänomen entstehen lassen 11
Abbildung 2: Wachstum der Datenmengen über die Zeit 12
Abbildung 3: Treiber für das Datenwachstum in deutschen Unternehmen 13
Abbildung 4: Erwarteter Business-Nutzen aus dem Einsatz von Big Data in deutschen Unternehmen 15
Abbildung 5: Herausforderungen bei der Planung und Umsetzung von Big-Data-Initiativen 16
Abbildung 6: Big Data in deutschen Unternehmen – Einsatz und Planungen 18
Abbildung 7: Beweggründe in deutschen Unternehmen für Beschäftigung mit Big Data 18
Abbildung 8: Merkmale von Big Data 19
Abbildung 9: Big Data als Ergänzung und Konkurrenz zur traditionellen IT 23
Abbildung 10: Analyseansätze für traditionelle Systeme und Big Data Systeme 26
Abbildung 11: Integrierte Anwendungslandschaft mit traditionellen Systemen und Big-Data-Lösungen 28
Abbildung 12: Globaler Markt für Big Data in drei Kennziffern 47
Abbildung 13: Entwicklung des globalen Big-Data-Marktes 2011-2016 48
Abbildung 14: Struktur des globalen Big-Data-Marktes 2011-2016 48
Abbildung 15: Struktur des globalen Big-Data-Marktes nach Regionen 2012 49
Abbildung 16: Deutscher Big-Data-Markt 2011-2016 nach Marktsegmenten 49
Abbildung 17: Ausgewählte Big-Data-Einsatzbeispiele in Kennziffern 52

1 Geleitwort



Prof. Dieter Kempf
BITKOM Präsident,
Vorsitzender des Vorstands Datev eG

Der im Januar 2012 gegründete BITKOM-Arbeitskreis Big Data stellt im vorliegenden Leitfaden seine ersten Arbeitsergebnisse zur Diskussion. Beleuchtet wird das Phänomen Big Data vorrangig in seiner wirtschaftlichen Dimension und mit Blick auf das Management von Unternehmen, denn Big-Data-Lösungen können der Wettbewerbsfähigkeit von Organisationen einen kräftigen Schub verleihen.

Wenn gelegentlich zu hören ist, die Informationswirtschaft würde mit Big Data lediglich den nächsten Hype entfachen, so sprechen die drei Dutzend Praxisbeispiele im Leitfaden eine andere Sprache.

Mit dem Leitfaden sollen Manager angeregt werden, sich mit dem neuen Thema vertieft auseinanderzusetzen: Bei Big Data geht es darum, die in bisher ungekanntem Umfang zur Verfügung stehenden, qualitativ vielfältigen und unterschiedlich strukturierten Daten in Geschäftsnutzen zu verwandeln. Im Leitfaden wird gezeigt, welche betrieblichen Funktionsbereiche von den neuen Möglichkeiten besonders profitieren können. Klar ist: Wir stehen hier noch ganz am Anfang!

Die größte Herausforderung wird darin liegen, dass Manager diese Möglichkeiten für ihre Entscheidungspraxis erschließen. Durch die schnellen Fortschritte vieler Technologien – etwa Cloud Computing, Social Media oder Mobile Computing – sind sie es gewohnt, sich permanent neuen Herausforderungen zu stellen. Big Data gehört zweifellos dazu.

Die Wirtschaft wird in naher Zukunft viele Big-Data-Spezialisten nachfragen. Für gestandene IT-Experten eröffnen sich in diesem Bereich interessante neue Betätigungsfelder. In den meisten Fällen werden die Unternehmen ihre Data Scientists unter den Hochschulabsolventen rekrutieren. An der TU Berlin und anderen Hochschulen werden die ersten Curricula angeboten. Wir plädieren für eine Ausweitung der Ausbildungsmöglichkeiten!

Big Data hat nicht nur wirtschaftliche Bedeutung. Wichtig ist zudem die politische Dimension. Wir dürfen uns keine Rahmenbedingungen auferlegen, die uns darin hindern, Big-Data-Methoden breit einzusetzen. Viele neue Fragen stehen im Raum: Wie sieht es mit dem intellektuellen Eigentum an den Ergebnissen von Big Data Analytics aus und wer trägt für sie Verantwortung? Ist es

möglich, den in Deutschland erreichten hohen Standard im Datenschutz in einen Standortvorteil im Bereich Big Data umzumünzen? Wie gelingt es uns, die in wichtigen Teilbereichen von Big Data erarbeiteten Positionen in der Forschung zügig in innovative, exportfähige Produkte und Services umzusetzen? Gemeint sind hier insbesondere die im THESEUS-Forschungsprogramm neu- und weiterentwickelten semantischen Technologien. Auch die öffentliche Verwaltung sollte die Einsatzmöglichkeiten von Big Data Analytics prüfen.

Das Autorenteam hat bereits die Arbeit am zweiten Leitfaden aufgenommen, dessen Schwerpunkt die konkrete Umsetzung von Big-Data-Projekten bilden wird. Wie werden solche Projekte erfolgreich aufgesetzt, wie die Mitarbeiter dazu befähigt? Welche Vorgehensmodelle versprechen Erfolg? Und welche Standards prägen den Markt?

Ich wünsche eine spannende Lektüre und freue mich auf den Dialog zu den aufgeworfenen Fragen!

Prof. Dieter Kempf, BITKOM-Präsident



2 Management Summary

Big Data bezeichnet die Analyse großer Datenmengen aus vielfältigen Quellen in hoher Geschwindigkeit mit dem Ziel, wirtschaftlichen Nutzen zu erzeugen.

Die Zusammenfassung des Leitfadens geht auf acht Punkte ein:

- In der digitalen Welt treten Daten als vierter Produktionsfaktor neben Kapital, Arbeitskraft und Rohstoffe.
- Viele Unternehmen werden konventionelle und neue Technologien kombinieren, um Big-Data-Lösungen für sich nutzbar zu machen.
- Der überwiegende Teil der in Unternehmen vorliegenden Daten ist unstrukturiert, kann aber in eine strukturierte Form überführt sowie quantitativen und qualitativen Analysen zugänglich gemacht werden.
- Empirische Studien sowie zahlreiche Einsatzbeispiele belegen den wirtschaftlichen Nutzen von Big Data in vielen Einsatzgebieten.
- Einige Funktionsbereiche von Unternehmen sind für den Big-Data-Einsatz prädestiniert. Dazu gehören Bereiche wie Marketing und Vertrieb, Forschung und Entwicklung, Produktion sowie Administration/Organisation/Operations.
- Der Einsatz von Big-Data-Methoden sollte bereits in der Konzeptionsphase aus rechtlicher Sicht geprüft werden.
- Hohe zweistellige Wachstumsraten im Markt über einen längeren Zeitraum sind ein weiterer Beleg für die wirtschaftliche Bedeutung von Big-Data-Lösungen.
- Es ist im volkswirtschaftlichen Interesse, Erfahrungen und Best Practices bei der Nutzung von Big Data effektiv zu kommunizieren.

Daten als Produktionsfaktor - Bedeutung und Nutzen von Big Data

Big Data bezeichnet die wirtschaftlich sinnvolle Gewinnung und Nutzung entscheidungsrelevanter Erkenntnisse aus qualitativ vielfältigen und unterschiedlich strukturierten Informationen, die einem schnellen Wandel unterliegen und in bisher ungekanntem Umfang[1] anfallen.

Big Data stellt Konzepte, Methoden, Technologien, IT-Architekturen sowie Tools zur Verfügung, um die geradezu exponentiell steigenden Volumina vielfältiger Informationen in besser fundierte und zeitnahe Management-Entscheidungen umzusetzen und so die Innovations- und Wettbewerbsfähigkeit von Unternehmen zu verbessern.

Die Entwicklung von Big Data wird als Anzeichen für einen Umbruch gesehen: Während sich die Bedeutung von Hardware und Software vermindert, nimmt die Bedeutung von Daten als Faktor der Wertschöpfung zu. In der digitalen Welt treten Daten als vierter Produktionsfaktor neben Kapital, Arbeitskraft und Rohstoffe in Erscheinung.

Big Data gewinnt zunehmend an Bedeutung, weil das Volumen der zur Verfügung stehenden Daten wie auch die Zahl der Datentypen wächst. Mit neuen Hard- und Software-basierten Verfahren lässt sich die Flut der meist unstrukturierten Daten nutzen. Big-Data-Analysen generieren erheblichen Mehrwert und beeinflussen die

[1] Facebook speichert über 100 Petabyte Daten in einem Hadoop-Disk-Cluster und verarbeitet täglich 500 Terabyte an Daten – nahezu in Echtzeit. »So werden täglich 2,7 Milliarden Likes verarbeitet, die bei Facebook sowie anderen Webseiten anfallen. Die 955 Millionen Nutzer des Sozialen Netzes teilen täglich 2,5 Milliarden Inhalte und laden 300 Millionen Fotos hoch. 70.000 Anfragen sind zu bearbeiten, die von Nutzern kommen oder automatisch generiert werden.« Vgl.: http://www.silicon.de/41571372/500-terabyte-taglich-datenrausch-bei-facebook/ (Abruf 27. August 2012)



Strukturen und die Ausrichtung von Organisationen und Unternehmen sowie das Management.[2]

Big Data setzt da ein, wo konventionelle Ansätze der Informationsverarbeitung an Grenzen stoßen, die Flut zeitkritischer Informationen für die Entscheidungsvorbereitung zu bewältigen.

Mit Big Data können Unternehmen schneller auf Marktveränderungen reagieren. Mit höherer Agilität steigern sie ihre Wettbewerbsfähigkeit. Bestehende Produkt- und Service-Angebote können in kürzerer Zeit verbessert und neue entwickelt werden.

Die mit Big Data verbundenen neuen Chancen entstehen nicht automatisch. Unternehmen müssen sich mit einigen Herausforderungen auseinandersetzen, die in erster Linie mit dem Management von Daten zusammenhängen. Manager müssen darauf achten, in den Unternehmen rechtzeitig die erforderlichen Kenntnisse zu entwickeln und die Entscheidungs- und Geschäftsprozesse anzupassen. Der Umgang mit Daten spiegelt die Unternehmenskultur wider – Qualität in diesem Bereich muss eine Angelegenheit aller Mitarbeiter sein.

Transformationsstrategie zur Nutzung von Big Data

Der rapide Anstieg des Datenvolumens ist u.a. auf Sensordaten, Kommunikation über Social-Media-Kanäle und Mobilkommunikation zurückzuführen. Big Data ist folglich im Kontext einer Vielzahl von Technologien und Disziplinen zu sehen, die dabei unterstützen, Unternehmen agil und innovativ zu machen.

Transaktionale Systeme, Data Warehouse, Business Intelligence, Dokumenten-Management- und Enterprise-Content-Management-Systeme sowie Semantik sind wichtige Technologien, die zu Big Data hinführen.

Mit Big Data werden Daten in unterschiedlichsten Formaten für Analysen zugänglich und für geschäftliche Prozesse und Entscheidungen nutzbar gemacht.

Für Unternehmen gilt es, die Herausforderungen der mit Big Data verbundenen Prozesse, Technologien und Qualifikationen zu meistern, ohne die bewährten Systeme und Prozesse zu gefährden oder vollständig neu zu entwickeln.

In vielen Unternehmen wird zukünftig eine Kombination von konventionellen und neuen Technologien zum Einsatz kommen, um Big-Data-Lösungen bereitzustellen. Die bereits in vielen Unternehmen vorhandenen Technologien erlauben einen graduellen Übergang auf eine Big-Data-Lösung oder eine Integration mit Big-Data-Techniken.

Nutzung unstrukturierter Daten in Big-Data-Verfahren

Der überwiegende Teil der in Unternehmen vorliegenden Daten ist unstrukturiert. Durch die inhaltliche Erschließung werden unstrukturierte Daten automatisiert in eine strukturierte Form überführt. So werden weitere quantitative und qualitative Analysen ermöglicht. Die Basis für Management-Entscheidungen wird erweitert, und Unternehmen können Vorteile im Wettbewerb erzielen.

Praxiseinsatz, wirtschaftlicher Nutzen, Innovationswelle

Der wirtschaftliche Nutzen von Big Data liegt für viele Einsatzgebiete klar auf der Hand. Das belegen empirische Studien[3] sowie die in diesem Leitfaden skizzierten Einsatzbeispiele[4].

Big Data wird als Vehikel immer bedeutsamer, die Herausforderungen im Umgang mit Massendaten zu meistern. Fünf Effekte treten in allen Wirtschaftszweigen ein:

---

[2] Dr. Ralf Schneider, CIO der Allianz Gruppe, prognostiziert, dass »Realtime Analytics in zehn Jahren die Spielregeln des gesamten Geschäfts verändert haben werden« und »Informationsverarbeitung und -analyse ist nicht mehr nur ein Business Enabler, sondern der Kern des Business selbst.«
Vgl.: Schneider, Ralf: »Neues Spiel mit Echtzeitanalyse«, in: Ellermann, Horst (Hrsg.): CIO Jahrbuch 2013 - Neue Prognosen zur Zukunft der IT. IDG Business Media GmbH, München, August 2012, S. 48-49.

- Erstens schafft Big Data Transparenz. Allein ein Mehr an Transparenz im »Datendschungel« hilft Unternehmen, den Überblick über die Geschäftsprozesse zu behalten und besser fundierte Entscheidungen zu treffen. Wenn allen Akteuren einer Organisation zeitnah die gleichen Informationen zur Verfügung stehen, kann das Innovationspotenziale heben und die Wertschöpfung erhöhen.

- Zweitens schafft Big Data Spielraum für erweiterte Simulationen. Performance-Daten in Echtzeit ermöglichen kontrollierte Experimente, um Bedürfnisse und Variabilität zu identifizieren und die Leistung zu steigern.

- Drittens verbessert Big Data den Kundenzugang. Big Data versetzt Organisationen in die Lage, feinkörnige Bevölkerungs- und Kundensegmente zu erstellen und ihre Waren und Dienstleistungen auf deren Bedarf zuzuschneiden. Eine detaillierte Segmentierung von Zielgruppen erleichtert deren Ansprache, vermindert die Streuverluste und somit auch die Kosten für Marketingkampagnen.

- Viertens unterstützt Big Data Entscheidungsprozesse. Die Analyse umfangreicher Daten in Echtzeit (»Embedded Analytics«) verbessert Entscheidungen als vollautomatischen Prozess oder als Entscheidungsgrundlage für das Management. Auf Algorithmen basierende Auswertungen riesiger Datenmengen können in allen Unternehmensbereichen zur Verminderung von Risiken und zur Verbesserung von Geschäftsprozessen beitragen und die Intuition von Entscheidern ergänzen.

- Fünftens schließlich lässt Big Data auch Chancen für neue Geschäftsmodelle, Produkte und Dienstleistungen entstehen. Absehbar ist auch die Entstehung von Geschäftsideen, die im Zusammenspiel mit dem Internet bestehende Unternehmen in ihrer Existenz bedrohen können.

Wirtschaftlicher Nutzen in ausgewählten Funktionsbereichen

Big Data wird eingesetzt, wo qualitativ unterschiedliche Daten in hohen Volumina anfallen. Dazu gehören Forschung und Entwicklung, Marketing und Vertrieb, Produktion, Distribution und Logistik sowie Finanz- und Risiko-Controlling. In diesen fünf Funktionsbereichen lässt sich der wirtschaftliche Nutzen von Big Data besonders eindrucksvoll und beispielhaft belegen.

- Big Data erleichtert es Marketing- und Vertriebsabteilungen, Produkt- und Service-Angebote zunehmend auf Kundensegmente oder einzelne Kunden zuzuschneiden und Streuverluste im Marketing zu vermindern.

- Ein hohes Potenzial für den Einsatz von Big Data schlummert in der Wissenschaft sowie in der betrieblichen Forschung und Entwicklung. In der Entwicklung der nächsten Produktgeneration helfen Social-Media-Analysen und die Auswertung von Sensordaten der gegenwärtig im Einsatz befindlichen Produkte.

- Mit dem Internet der Dinge oder Machine-to-Machine-Kommunikation können produzierende Unternehmen ihre Fertigungsprozesse optimieren. Dafür erfassen Sensoren an Produkten und entlang von Produktions- und Lieferketten Daten – auch im späteren Betrieb. Viele Unternehmen arbeiten daran, die verschiedenen Unternehmensbereiche zu

---

[3] Vgl. u.a.: Global Survey: The Business Impact of Big Data. Avanade, November 2010. | Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, June 2011. | Big Data Analytics. TDWI Best Practice Report. Philip Russom, Fourth Quarter 2011, tdwi.org. | Big Data. Eine Marktanalyse der Computerwoche, München, November 2011. | Global Survey: Is Big Data Producing Big Returns? Avanade, June 2012 | Big Data Analytics in Deutschland 2012. White Paper, IDC, Mathias Zacher, Januar 2012. | Datenexplosion in der Unternehmens-IT: Wie Big Data das Business und die IT verändert (Eine Studie der Experton Group AG im Auftrag der BT (Germany) GmbH & Co. oHG), Dr. Carlo Velten, Steve Janata, Mai 2012 | Quo vadis Big Data. Herausforderungen – Erfahrungen – Lösungsansätze. TNS Infratest, August 2012 | Lünendonk®-Marktstichprobe 2012 »Business Intelligence als Kernkompetenz«, 22. August 2012

[4] vgl. Kapitel 10

verknüpfen und in die Optimierung auch Zulieferer und Partner einzubinden.

- In Distribution und Logistik geht es um nachhaltige Kostensenkung auf dem Wege einer stärkeren Vernetzung von Fahrzeugen mit der Außenwelt. Immer mehr Fahrzeuge werden mit Sensoren und Steuerungsmodulen ausgestattet, die Fahrzeugdaten wie den Benzinverbrauch, den Zustand von Verschleißteilen oder Positionsdaten erfassen und in Datenbanken übertragen. Mit diesen Daten können Disponenten zeitnah Transporte planen, gegebenenfalls Routen und Beladung ändern, Wartungskosten und Stillstandzeiten minimieren.

- Das Finanz- und Risiko-Controlling profitiert u.a. von neuen Möglichkeiten im Bereich Betrugserkennung und Risikomanagement. Bei der Betrugserkennung steht in erster Linie eine möglichst vollständige Sammlung und Beobachtung relevanter Handlungen. Das Risikomanagement wird durch hochkomplexe Berechnungen verfeinert.

Rechtsfragen

In Deutschland werden die Risiken von Big Data betont und Befürchtungen vor unkontrollierter Überwachung thematisiert. Bei allen berechtigten und notwendigen Hinweisen auf auftretende Risiken sollte das Augenmerk darauf gerichtet werden, die großen Chancen von Big Data zielgerichtet zu erschließen. Dafür existieren auch die rechtlichen Grundlagen, denn Big-Data-Methoden sind nach deutschem Datenschutzrecht in einer ganzen Reihe von Fällen zulässig.

Die rechtlichen Herausforderungen bestehen darin, in Vertragsverhältnissen zu beurteilen, welche Datenverarbeitung erforderlich ist, für wirksame Einwilligungen zu sorgen und taugliche Verfahren zum Privacy-Preserving Data Mining anzuwenden. Vor allem ist es wichtig, die rechtliche Zulässigkeit bereits bei der Entwicklung einer Big-Data-Anwendung zu prüfen. Die rechtliche Zulässigkeit hängt nämlich stark vom Design des Verfahrens ab. In der Anfangsphase der Entwicklung lässt sich das einfacher ändern als später, wenn ein Verfahren bereits eingeführt ist.

Markt für Big-Data-Lösungen

Big Data steht trotz der frühen Marktphase schon für ein hoch relevantes IT-Marktsegment. Die globalen Umsätze im Segment Big Data lagen 2011 in der Größenordnung von 3,3 Milliarden Euro. 2012 wird mit 4,5 Milliarden Euro gerechnet, und 2016 wird der globale Big-Data-Markt 15,7 Milliarden Euro schwer sein. Das entspricht einer mittleren Wachstumsrate von 36%.

Deutsche Unternehmen steigen bedächtig in das neue Thema ein, werden sich aber zügig zu »Big Data Champions« profilieren, um ihre Produktions-, Logistik- und Vertriebsketten weltweit optimiert zu planen und zu steuern. Für eine hoch wettbewerbsfähige und exportorientierte Volkswirtschaft führt an Big Data kein Weg vorbei.

Einsatzbeispiele von Big Data in Wirtschaft und Verwaltung

Der Mangel an Anwendungsbeispielen gilt als Hürde für den Erfolg am Markt. Die in diesem Leitfaden präsentierten drei Dutzend Einsatzbeispiele belegen die wirtschaftliche Relevanz von Big Data eindrucksvoll: Einige Beispiele mussten anonymisiert werden. Volkswirtschaftlich besteht ein Interesse, Best Practices zügig breitenwirksam werden zu lassen. Entscheider stehen vor der Herausforderung, auf kreative Weise die Chancen von Big Data für das Business auszuleuchten.

3 Big Data – Chancen und Herausforderungen für Unternehmen

Der Einsatz von Big Data in Unternehmen und Organisationen hat bereits begonnen. Big Data stellt Konzepte, Technologien und Methoden zur Verfügung, um die geradezu exponentiell steigenden Volumina vielfältiger Informationen noch besser als fundierte und zeitnahe Entscheidungsgrundlage verwenden zu können und so Innovations- und Wettbewerbsfähigkeit von Unternehmen weiter zu steigern.

Big Data setzt da an, wo konventionelle Ansätze der Informationsverarbeitung an Grenzen stoßen, die Flut und Komplexität zeitkritischer Informationen für die Entscheidungsvorbereitung zu bewältigen.

Big Data versetzt Unternehmen in die Lage, bei der Vorbereitung von Management-Entscheidungen in neue Dimensionen vorzustoßen. Die Business Performance in unterschiedlichsten Geschäftsbereichen kann noch besser hinterfragt, mögliche Optimierungspotenziale identifiziert und die damit notwendigen Massnahmen abgeleitet werden. Auf Veränderungen in den Märkten kann somit schneller reagiert werden. Höhere Agilität verbessert die Wettbewerbsfähigkeit. Bestehende Angebote können in kürzerer Zeit verbessert und noch zielgerichteter den Kundenwünschen angepasst werden. Neueste Informationen über Märkte, Trends und Kundenbedürfnisse sind damit die Grundlage für die Entwicklung völlig neuer Services.

Die mit Big Data verbundenen neuen Chancen entstehen nicht automatisch. Unternehmen müssen sich mit Herausforderungen auseinandersetzen, die primär mit dem Management von Daten zusammenhängen.

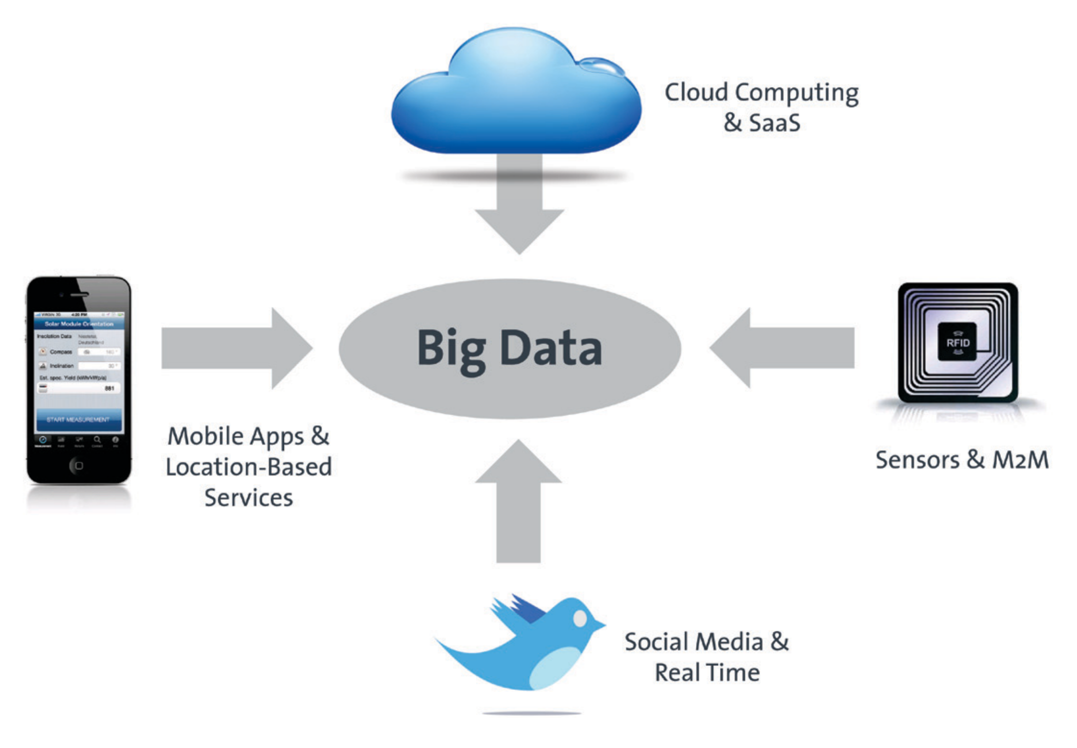


Abbildung 1: Welche Informationstechnologien das Big-Data-Phänomen entstehen lassen; Quelle: Experton Group 2012

Fernsehen, Computer, PC und Internet - schon die zweite Hälfte des vergangenen Jahrhunderts galt als Informationszeitalter. Trotzdem war die erzeugte Informationsmenge damals gering im Vergleich mit der Datenexplosion, die in der letzten Dekade stattgefunden hat. Technologien wie RFID, Ambient Intelligence, Smartphones und die immer stärkere Akzeptanz und Nutzung von Social-Media-Anwendungen wie Blogs und Foren oder Facebook und Twitter (vgl. Abbildung 1) lassen das Datenaufkommen explodieren (vgl. Abbildung 2).

Jahrzehnten entsprechend dem Mooreschen Gesetz etwa alle 18 Monate verdoppeln.[6] Im Jahr 2020 wird das weltweite Datenvolumen dann über 100 Zettabytes erreichen: 100.000.000.000.000.000.000.000 Byte sind eine wahrhaft unvorstellbare Menge.

Die Datenmenge an sich ist aber nur eine statistische Größe. Mit der Datenexplosion geht eine Zunahme der Datenvielfalt einher. Die Daten stammen zum Beispiel aus Internet-Transaktionen, Social Networking, Machine-

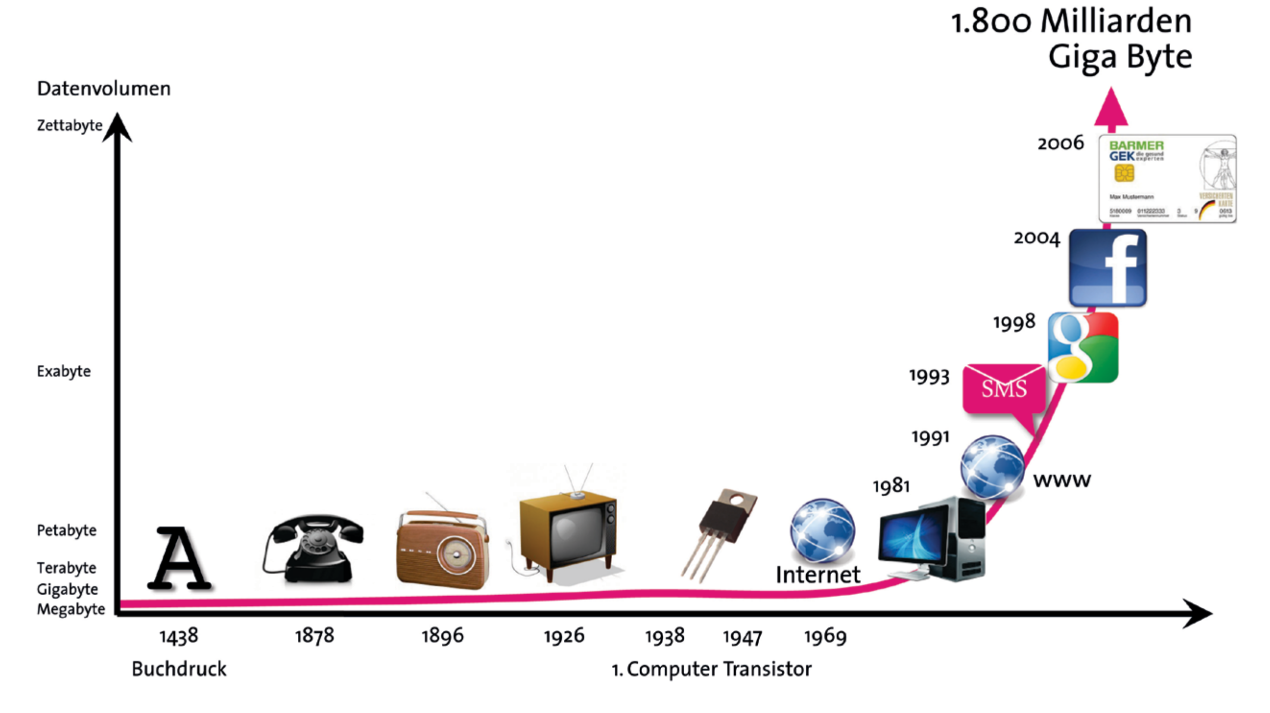


Abbildung 2: Wachstum der Datenmengen über die Zeit

Von 2000 bis 2002 sind mehr Daten generiert worden, als in den 40.000 Jahren davor. Von 2003 bis 2005 hat sich diese Datenmenge wiederum vervierfacht. Und 2012 wird sich das weltweite Volumen digitaler Daten auf 2.5 Zettabytes gegenüber 2006 verzehnfachen.[5] Es ist davon auszugehen, dass das weltweite Datenvolumen in den nächsten Jahren schneller wachsen wird als die Kapazitäten zur Datenverarbeitung, die sich seit einigen

to-Machine-Lösungen, Sensoren, mobilen Endgeräten oder Messungen jeglicher Art (vgl. Abbildung 3). Dabei ist zu beachten, dass rund 85 Prozent der Daten unstrukturiert sind, aber wertvolle Informationen enthalten. Diese konnten aber bisher kaum mit vertretbarem Aufwand ausgewertet und für die Entscheidungsvorbereitung genutzt werden.

[5] Vgl. Forrester: »Measure And Manage Brand Health«, July 30, 2012, Research Report
[6] Vgl. http://blogs.forrester.com/holger_kisker/12-08-15-big_data_meets_cloud (Abruf 21. August 2012)

Welche Treiber sind für das Datenwachstum in Ihrem Unternehmen wesentlich?

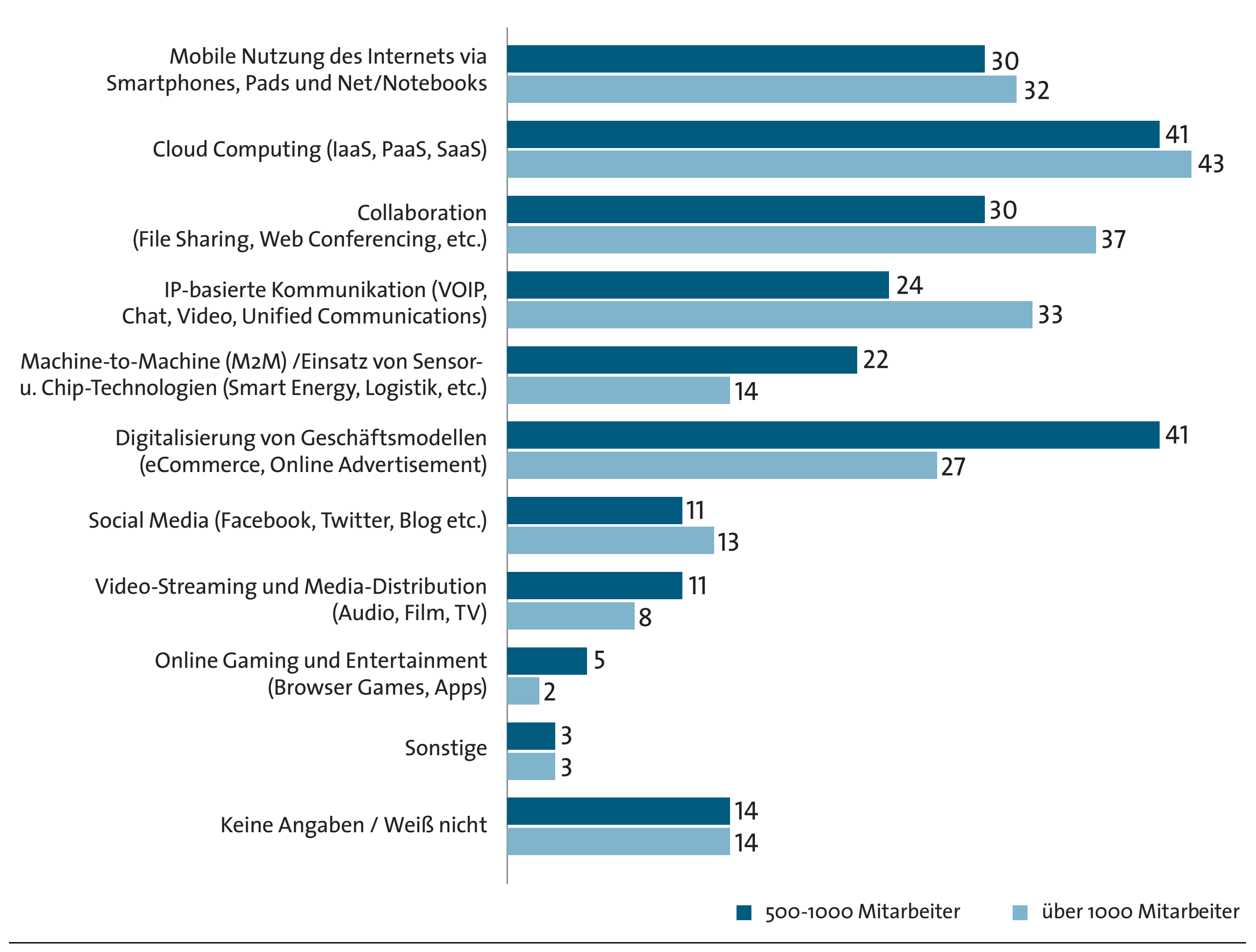


Abbildung 3: Treiber für das Datenwachstum in deutschen Unternehmen; Quelle: Experton Group 2012



3.1 Mit Big Data verbundene Chancen

Erfreulicherweise sind parallel mit der rasanten Explosion der Datenmengen auch die technischen Möglichkeiten gewachsen, Informationen in großen Mengen und hoher Geschwindigkeit zu verarbeiten und zu analysieren. Hieraus eröffnen sich vielfältige Chancen, um Unternehmen aufgrund neuer und bislang unzugänglicher Daten und Informationen besser auf Marktanforderungen einzustellen (vgl. Tabelle 1). Diese Chancen werden auch im Management klar gesehen (vgl. Abbildung 4).

Aspekt	Erläuterung
Big-Data-Strategie	Enorm zunehmende Datenvolumina liefern Argumente für die Bedeutung der Daten und regen das Management an, eine strategische Antwort für die erfolgreiche Nutzung neuer Daten sowie verfügbarer Technologien zu entwickeln.
Governance	Durch die Festlegung klarer Verantwortlichkeiten ist Transparenz zu schaffen. Ein zeitnaher Zugang zu Big Data für alle Beteiligten erschließt Innovationspotenzial und fördert neue Wertschöpfung.
Prozessoptimierung	Fortgeschrittenes Datenmanagement bildet die Voraussetzung für die Prozessoptimierung. Die Bereitstellung von Informationen in Echtzeit ist die Grundlage für völlig neue Geschäftsprozesse.
Compliance	Das wachsende Datenvolumen sowie die Komplexität der Daten stellen hohe Anforderungen an die Compliance. Ein transparentes Datenmanagement einhergehend mit der detaillierten und zeitnahen Aufbereitung vorgeschriebener Informationen erleichtert die Erfüllung regulatorischer Anforderungen.
Kundenkenntnis	Big Data versetzt Unternehmen in die Lage, Kundensegmente mit größerer Granularität im Auge zu behalten und somit Produkt- und Serviceangebote besser am realen Bedarf auszurichten.
Angemessene Entscheidungsbasis	Neue Formen der Datenauswertung und -aggregation liefern Informationen mit bisher nicht erreichter Spezifik. Die Analyse der explodierenden Datenmengen erleichtert fundierte Management-Entscheidungen zum bestmöglichen Zeitpunkt. Die Analyse komplexer Datensätze in Verbindung mit problemadäquater Interpretation der Ergebnisse verbessert Entscheidungen als zunehmend automatisierter Prozess, oft in Echtzeit (»Embedded Analytics«) Die Verknüpfung unterschiedlicher Datenformate unterstützt völlig neue Erkenntnisse.
Time-to-Market	Mit dem frühzeitigen Erkennen von Marktveränderungen wird Reagieren zunehmend durch Agieren ersetzt.
Geschäftsmodelle	Big Data lässt Unternehmen nicht nur bestehende Angebote verbessern, sondern völlig neue Angebote entwickeln.
Vereinfachung der Systeminfrastruktur	Die wachsenden technischen Möglichkeiten bieten Unternehmen die Chance, mit ihren Big-Data-Projekten die Systeminfrastruktur zu vereinfachen und ggf. die Anzahl der Datenbanken bzw. Data-Warehouse-Installationen zu reduzieren.

Tabelle 1: Chancen durch Big Data



Abbildung 4: Erwarteter Business-Nutzen aus dem Einsatz von Big Data in deutschen Unternehmen; Quelle: Experton Group 2012

3.2 Mit Big Data verbundene Herausforderungen

Die schiere Menge an Informationen, deren Vielfalt sowie deren immer kürzere Aktualität stellt Unternehmen vor große Herausforderungen (vgl. Abbildung 5 und Tabelle 2), die allerdings auch neue Chancen bieten. Unternehmen müssen Antworten für diese Fragestellungen entwickeln.

Gefragt sind innovative Big-Data-Strategien. Ohne eine solche Strategie droht die Gefahr, dass Unternehmen im Datendschungel versinken oder auf wesentliche Markt- und Kundenveränderungen zu spät reagieren. Mit einer Big-Data-Strategie legen Unternehmen das Fundament, bevorstehende Veränderungen frühzeitig zu erkennen und sich dafür optimal aufzustellen. Wettbewerbsvorteile können erlangt werden.

Immer größere Datenmengen allein sind allerdings keine Garantie, die Chancen von Big Data zu erschließen. Denn schließlich verfügen Unternehmen schon seit Jahren über Massendaten und Dokumente, die sie mithilfe von Business Intelligence oder Data-Warehousing auswerten und bereitstellen. Allerdings stoßen die bisherigen Verfahren angesichts der Datenflut an ihre Grenzen: Die Auswertung dauert zu lange und verliert für die Entscheidungsvorbereitung an Wert.

Einer Umfrage[7] unter mehr als 500 Managern und IT-Entscheidern aus 17 Ländern zufolge ist ein Großteil vom Datenaufkommen am Arbeitsplatz überwältigt. In dieser Flut sehen sich viele Manager nicht in der Lage, Entscheidungen rechtzeitig zu treffen, obwohl mehr als zwei Drittel der Befragten glauben, die richtigen Daten seien vorhanden. Allerdings wusste ein Drittel nicht, wen im Unternehmen sie auf der Suche nach den richtigen Informationen fragen könnten.

7 Vgl.: Avanade (November 2010): Global Survey: The Business Impact of Big Data. Download von http://www.avanade.com/Documents/Research%20and%20Insights/Big%20Data%20Executive%20Summary%20FINAL%20SEOv.pdf (am 10. August 2012).
Vgl auch: http://www.avanade.com/de-de/about/avanade-news/press-releases/Pages/Globale-Avanade-Studie-Jedes-dritte-deutsche-Unternehmen-von-Datenfluss-%C3%BCberfordert-page.aspx sowie: Avanade (June 2012): Global Survey: Is Big Data Producing Big Returns?

Wer Big-Data-Lösungen einsetzen will, muss also einige Herausforderungen meistern. So ist neben neuen Hard- und Software-Infrastrukturen eine klare Strategie und Roadmap erforderlich, mit der sich Unternehmen Schritt für Schritt Big Data annähern können. Dazu gehört es vor allem, Transparenz im Datenbestand, in den Datenquellen und in der Datenvielfalt herzustellen, um Daten überhaupt effektiv managen, validieren und analysieren zu können. Wer nicht weiß, welche Informationen in welcher Form überhaupt vorhanden sind, wird scheitern.

Informationssicherheit, Datenschutz, Data Governance

Auch die Informationssicherheit und der Datenschutz unterliegen ganz neuen Anforderungen. Zu einer der ersten Maßnahmen vor der Umsetzung von Big Data Daher gehört daher der Aufbau einer Data Governance, die Prozesse und Verantwortlichkeiten festlegt und Compliance-Richtlinien definiert. Schließlich darf nicht außer Acht gelassen werden, dass die Total Cost of Ownership steigen, insbesondere dann, wenn die Auswertung komplexer Daten in Echtzeit erfolgen soll. Diese Herausforderungen führen dazu, dass klare Verantwortlichkeiten und Prozesse im Umgang mit Big Data definiert werden müssen.

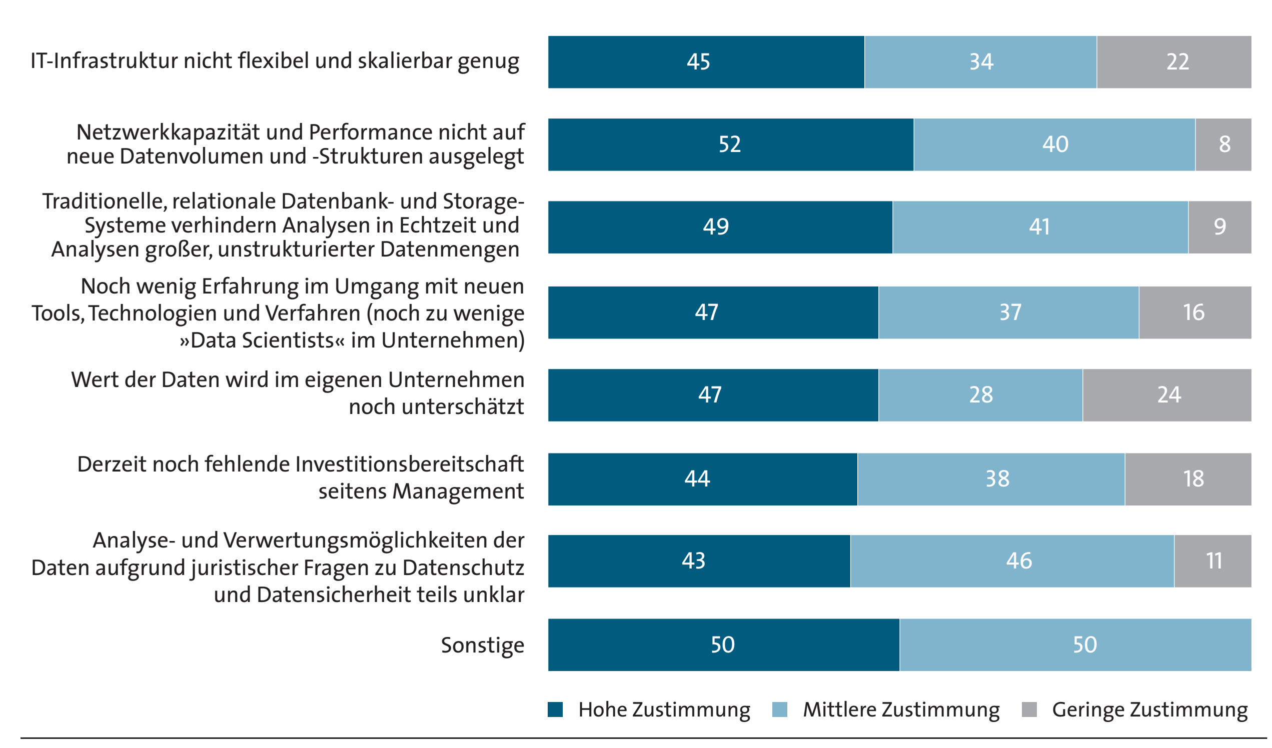


Abbildung 5: Herausforderungen bei der Planung und Umsetzung von Big-Data-Initiativen; Quelle: Experton Group 2012



Herausforderung	Erläuterung
Total Cost of Ownership	Heterogene und komplexe Informationslandschaften führen zu steigenden Betriebskosten.
Datenverluste	Unternehmen müssen sicherstellen, dass sie alle wesentlichen Daten und Informationen erfassen und diese nicht verlorengehen. Durch schnell wachsende Volumina stellt sich diese Aufgabe in einer neuen Dimension.
IT-Sicherheit, Betrugs- und Manipulations-Prävention	Volumen, Komplexität und Wert von Informationen verstärken die Notwendigkeit, Betrug und Manipulation vorzubeugen.
Transparenz	Das immense Volumen sowie die Vielfalt der Informationen erhöhen die Bedeutung geeigneter Datenstrukturen. Klare Datenstrukturen und Abläufe sind aber die Grundlage, um die Interpretation der Daten, die Informationsflüsse, die Verantwortlichkeiten und damit die Transparenz über das Datenuniversum zu ermöglichen
Dateninterpretation, Validierung	Das immense Volumen sowie die Vielfalt der Informationen erhöhen die Anforderungen an die problemadäquate Interpretation der Daten sowie die Sicherstellung ihrer Aktualität. Die Früherkennung von Signalen wichtiger Veränderungen nimmt an Bedeutung zu. Gleichzeitig gilt es, Fehlinterpretationen zu verhindern.
Entscheidungsbasis	Hohes Datenvolumen und zunehmend volatile Märkte erfordern und erschweren gleichzeitig eine schnelle und akkurate Datenanalyse als Basis für Management-Entscheidungen.

Tabelle 2: Mit Big Data verbundene Herausforderungen an Unternehmen

Unternehmen, die diese Herausforderungen als Chance verstehen, sie hinterfragen und entsprechende Strategien und Lösungsansätze ableiten, können sich dagegen agil auf neue Kunden- und Marktsituationen einstellen sowie Möglichkeiten für Geschäftsoptimierungen wahrnehmen, die zu einem signifikanten Wettbewerbsvorteil führen können.

Die Abbildung 6 verdeutlicht, dass deutsche Unternehmen bei strategischen Überlegungen zu Big Data noch am Anfang stehen. Handfeste Gründe (vgl. Abbildung 7) sprechen jedoch dafür, dass in nächster Zeit deutlich mehr Unternehmen die Option Big Data genau prüfen werden.



Hat sich Ihr Unternehmen bereit mit Big Data beschäftigt?

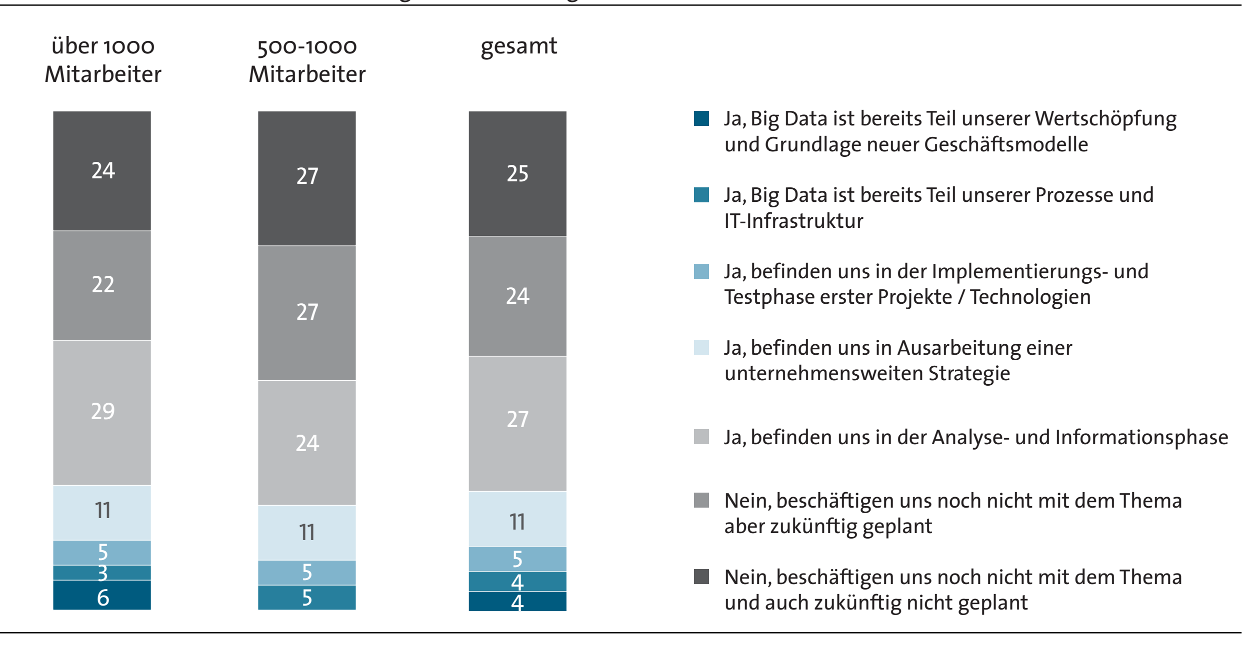


Abbildung 6: Big Data in deutschen Unternehmen – Einsatz und Planungen; Quelle: Experton Group 2012

Welche Beweggründe hat Ihr Unternehmen bei der Beschäftigung mit Big Data?



Abbildung 7: Beweggründe in deutschen Unternehmen für Beschäftigung mit Big Data; Quelle: Experton Group 2012



4 Big Data – Begriffsbestimmung

Big Data unterstützt die wirtschaftlich sinnvolle Gewinnung und Nutzung entscheidungsrelevanter Erkenntnisse aus qualitativ vielfältigen und unterschiedlich strukturierten Informationen, die einem schnellen Wandel unterliegen und in bisher ungekanntem Umfang zu Verfügung stehen. Big Data spiegelt den technischen Fortschritt der letzten Jahre wider und umfasst dafür entwickelte strategische Ansätze sowie eingesetzte Technologien, IT-Architekturen, Methoden und Verfahren.

Mit Big Data erhalten Manager eine deutlich verbesserte Grundlage für die Vorbereitung zeitkritischer Entscheidungen mit besonderer Komplexität.

Aus Business-Perspektive verdeutlicht Big Data, wie auf lange Sicht die Daten zu einem Produkt werden. Big Data öffnet die Perspektive auf die »industrielle Revolution der Daten«, während gleichzeitig Cloud Computing den IT-Betrieb industrialisiert.

Aus IT-Perspektive markiert Big Data die aufkommenden Herausforderungen sowie neuen technologischen Möglichkeiten für Speicherung, Analyse und Processing schnell wachsender Datenmengen.

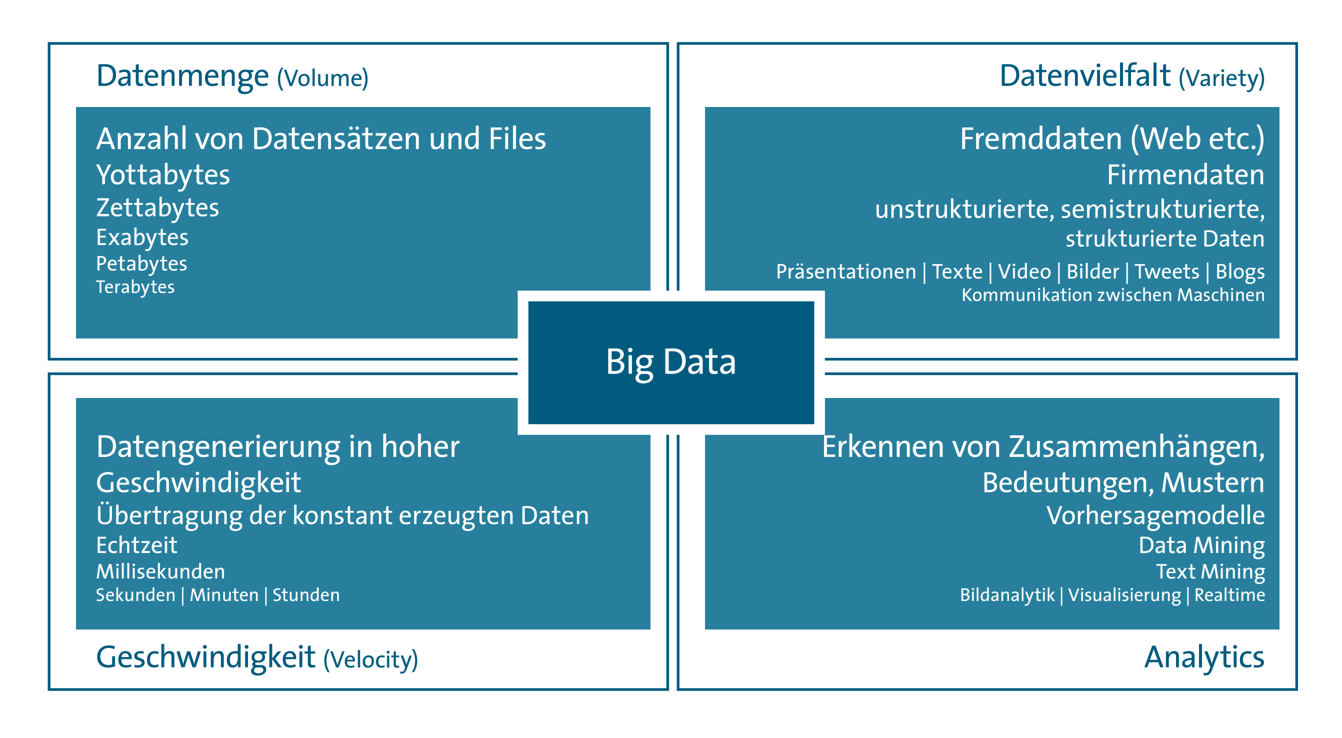


Abbildung 8: Merkmale von Big Data

Seit vielen Jahren kämpfen Unternehmen und Organisationen mit schnell wachsenden Datenbergen. Informatiker entwickelten zahlreiche Ansätze, die Probleme zu mildern. Die Diskussionen wurden jedoch vorrangig in den Spezialistenkreisen geführt. Das änderte sich vor drei-vier Jahren schlagartig, nachdem einige Publikationen Big Data in das Bewusstsein von Entscheidungsträgern gerückt hatten:

- Im September 2008 platzierte das Wissenschaftsmagazin »Nature« den Begriff »Big Data« auf der Titelseite[8] einer Sonderausgabe und verwies auf ein besonderes Problem: In vielen Bereichen der Forschung überstiegen die in Experimenten und Simulationen erzeugten Datenmengen alle zuvor gekannten und beherrschbaren Größenordnungen.
- Im Februar 2010 beschrieb »The Economist« die große »Datenflut« in ihren wirtschaftlichen Dimensionen und charakterisierte die inhärenten Chancen und Herausforderungen.[9]
- Im Mai 2011 publizierte das Analystenhaus Forrester Research eine Referenz-Architektur für Big-Data-Anwendungen[10].
- Im Juni 2011 schließlich erschien die erste große Studie, die sich ausschließlich mit dem wirtschaftlichen Potential von »Big Data« beschäftigte[11]. Im Juli 2011 nahm das Analystenhaus Gartner den Begriff »Big Data« in seinen »Hype Cycle« zur Bewertung neuer Technologien auf.[12]

So hat ein schon länger beobachtetes Phänomen einen prägnanten Namen erhalten.

Zahlreiche Marktteilnehmer haben aus unterschiedlichen Perspektiven Definitionen für »Big Data« vorgeschlagen. Mittlerweile zeichnet sich ein Konsens in der grundsätzlichen Beschreibung ab.

Es handelt sich im Wesentlichen um ein technologisches Phänomen im Bereich Datenmanagement und Datenanalyse, dem eine große Bedeutung für die Weiterentwicklung von Geschäftsprozessen quer durch alle Branchen und Unternehmensfunktionen beigemessen wird.

Big Data weist vier wesentliche Facetten auf (vgl. Tabelle 3 und Abbildung 8).

8 Vgl. http://www.nature.com/nature/journal/v455/n7209/covers/ (Abruf 26.07.2012)
9 Vgl.: http://www.economist.com/node/15579717 (Abruf 26.07.2012)
10 Forrester Research (May 2011): «Big Opportunities in Big Data" Report. http://www.forrester.com/home#/Big+Opportunities+In+Big+Data/fulltext/-/E-RES59321 (Abruf 26.07.2012))
11 McKinsey Global Institute (June 2011): »Big data: The next frontier for innovation, competition, and productivity«. Vgl.: http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation (Abruf 26.07.2012)
12 Vgl. http://www.gartner.com/it/page.jsp?id=1763814 (Abruf 27.08.2012)

Facette	Erläuterung
Datenmenge (Volume)	Immer mehr Organisationen und Unternehmen verfügen über gigantische Datenberge, die von einigen Terabytes bis hin zu Größenordnungen von Petabytes führen. Unternehmen sind oft mit einer riesigen Zahl von Datensätzen, Dateien und Messdaten konfrontiert.
Datenvielfalt (Variety)	Unternehmen haben sich mit einer zunehmenden Vielfalt von Datenquellen und Datenformaten auseinanderzusetzen. Aus immer mehr Quellen liegen Daten unterschiedlicher Art vor, die sich grob in unstrukturierte[13], semistrukturierte[14] und strukturierte[15] Daten gruppieren lassen. Gelegentlich wird auch von polystrukturierten Daten gesprochen. Die unternehmensinternen Daten werden zunehmend durch externe Daten ergänzt, beispielsweise aus sozialen Netzwerken. Bei den externen Daten sind z. B. Autoren oder Wahrheitsgehalt nicht immer klar[16], was zu ungenauen Ergebnissen bei der Datenanalyse führen kann.
Geschwindigkeit (Velocity)	Riesige Datenmengen müssen immer schneller ausgewertet werden, nicht selten in Echtzeit. Die Verarbeitungsgeschwindigkeit hat mit dem Datenwachstum Schritt zu halten. Damit sind folgende Herausforderungen verbunden: Analysen großer Datenmengen mit Antworten im Sekundenbereich, Datenverarbeitung in Echtzeit, Datengenerierung und Übertragung in hoher Geschwindigkeit.
Analytics	Analytics umfasst die Methoden zur möglichst automatisierten Erkennung und Nutzung von Mustern, Zusammenhängen und Bedeutungen. Zum Einsatz kommen u.a. statistische Verfahren, Vorhersagemodelle, Optimierungsalgorithmen, Data Mining, Text- und Bildanalytik. Bisherige Datenanalyse-Verfahren werden dadurch erheblich erweitert. Im Vordergrund stehen die Geschwindigkeit der Analyse (Realtime, Near-Realtime) und gleichzeitig die einfache Anwendbarkeit, ein ausschlaggebender Faktor beim Einsatz von analytischen Methoden in vielen Unternehmensbereichen.

Tabelle 3: Facetten von Big Data

Zusammenfassend bezeichnet Big Data den Einsatz großer Datenmengen aus vielfältigen Quellen mit einer hohen Verarbeitungsgeschwindigkeit zur Erzeugung wirtschaftlichen Nutzens.

»Big Data« liegt immer dann vor, wenn eine vorhandene Unternehmensinfrastruktur nicht mehr in der Lage ist, diese Datenmengen und Datenarten in der nötigen Zeit zu verarbeiten. Zur Kennzeichnung dieser Verarbeitung setzt sich der Begriff »Analytics« durch. Es liegt nahe, Big Data schnell und gründlich zu analysieren, um möglichst viel Wert aus den Datenbergen zu schöpfen.

Entscheidend für die neue Qualität ist, dass herkömmliche Technologien zur Datenverarbeitung nicht mehr ausreichen und deshalb spezielle Big-Data-Technologien zum Einsatz kommen.

Dabei ist Big Data ein ambivalentes Phänomen. Zum einen stehen Unternehmen und Organisationen vor einer

Big Data bezeichnet den Einsatz großer Datenmengen aus vielfältigen Quellen mit einer hohen Verarbeitungsgeschwindigkeit zur Erzeugung wirtschaftlichen Nutzens.

Reihe von technischen und organisatorischen Herausforderungen, um mit der Datenflut zurechtzukommen. Zum anderen sind Daten durchaus gewollte Ergebnisse einer immer genaueren Beobachtung von Geschäftsprozessen – und damit eine große Chance, Wettbewerbsvorteile zu erzielen.

Möglich wird dies durch die Weiterentwicklung in Bereichen wie Sensorik[17] oder Social Media[18]. Die mit Big Data verbundene Erwartung ist, Geschäftsprozesse besser zu verstehen und letztlich wirksamer zu managen.

13 z. B. Präsentationen, Texte, Video, Bilder, Tweets, Blogs
14 z. B. Kommunikation von und zwischen Maschinen
15 z. B. von transaktionalen Applikationen
16 Verstärken kann sich dieser Effekt bei unstrukturierten Daten, wie z. B. bei der Bildauswertung.
17 RFID, Bewegungssensoren in Smartphones, intelligente Stromzähler…
18 Facebook-Kommentare als Indikatoren für Meinungen und Bewertungen.

5 Einordnung von Big Data in die Entwicklungslinien der Technologien und Transformationsstrategien

Wie steht Big Data zu vorhandenen Technologien? Handelt es sich dabei um eine logische Weiterentwicklung bestehender Technologien oder einen technologischen Umbruch? Die Entwicklung von Big Data lässt sich als ein Anzeichen für einen weiteren fundamentalen Umbruch ansehen: Hardware und Software treten als Faktor der Wertschöpfung zurück und die Daten werden zum entscheidenden Erfolgsfaktor.

Eine Einordnung von Big Data in die Entwicklungslinien der Technologien zeigt verschiedene Anknüpfungspunkte an bestehende Technologien, z. B. transaktionale Systeme, Data Warehouse, Business Intelligence, Dokumenten-Management- und Enterprise-Content-Management-Systeme.

Die inhaltliche Auseinandersetzung mit Daten und ihrer Semantik gewinnen mit Big Data an Bedeutung: Big Data verspricht, Daten in unterschiedlichsten Formaten zugänglich und für geschäftliche Prozesse und Entscheidungen nutzbar zu machen. Dafür werden unzugängliche Daten mit Hilfe von semantischen Technologien in »ausreichend« strukturierte Daten »übersetzt«.

Für Unternehmen gilt es, die Herausforderungen der mit Big Data verbundenen Prozesse, Technologien und Qualifikationen zu meistern, ohne die bewährten Systeme und Prozesse zu gefährden oder vollständig neu zu entwickeln.

In Unternehmen wird zukünftig eine Kombination von konventionellen und neuen Technologien zum Einsatz kommen, um Big-Data-Lösungen bereitzustellen. Die bereits in vielen Unternehmen vorhandenen Technologien erlauben einen graduellen Übergang auf eine Big-Data-Lösung oder eine Integration mit Big-Data-Techniken.

Immer noch verdoppeln sich die erreichbare Speichergröße und Verarbeitungsgeschwindigkeit spätestens alle zwei Jahre und folgen damit seit mehr als vier Jahrzehnten dem Mooreschen Gesetz. Große Datenmengen - Big Data - gibt es schon immer in den existierenden IT-Landschaften: Datenvolumen an der Grenze des technisch Machbaren. Big Data kann als logische Weiterentwicklung bestehender Technologien gesehen werden, genauso als eine direkte Konkurrenz zu den etablierten Systemen oder als sinnvolle Ergänzung einer Anwendungs-Landschaft (vgl. Abbildung 9).

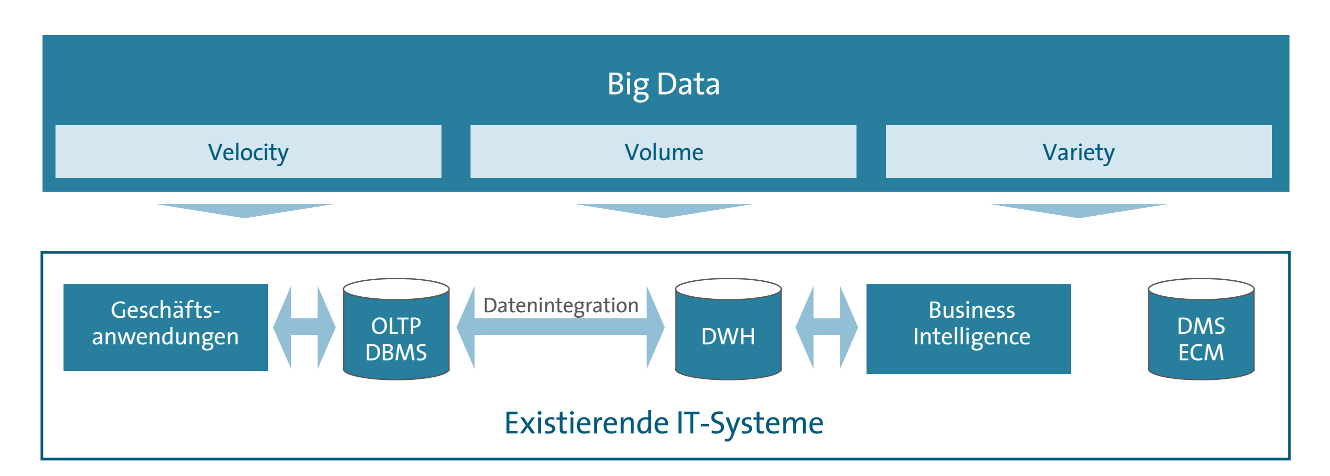


Abbildung 9: Big Data als Ergänzung und Konkurrenz zur traditionellen IT

5.1 Transaktionale Systeme

Wachsende Datenmengen, die durch die Digitalisierung der Geschäftswelt vorangetrieben werden[19], sind in der Vergangenheit immer vor allem durch transaktionale[20] und analytische[21] Systeme bedient worden. Derartige Systeme finden sich in allen Firmen; es existiert ein breiter und etablierter Markt mit einsatzfähigen Software-Lösungen. Dabei sind transaktionale Geschäftsanwendungen und operative Datenbanken heute vor allem darauf optimiert, hochwertige Geschäftsdaten bei einer möglichst hohen Datenkonsistenz und einem hohen Transaktionsdurchsatz zu verarbeiten – immer auf Basis von strukturierten Daten[22]. Steigende transaktionale Datenmengen und der zunehmende Bedarf von Fachabteilungen und Management an umfangreichen Ad-hoc-Analysen und -Berichten führen zu einem explosionsartigen Wachstum der technischen Anforderungen an die zugrundeliegenden Systeme. Selbst bei der Lösung der gleichen Aufgaben unterscheiden sich die Schwerpunkte zwischen Big-Data-Lösungen und transaktionalen und analytischen Systemen deutlich (vgl. Tabelle 4).

Transaktionale Systeme	Big Data
Basiert auf strukturierten und gut dokumentierten Daten (z. B. relationalen Tabellenmodellen)	Starke Betrachtung von unstrukturierten oder polystrukturierten Daten
Integrität durch das Datenmodell sicher gestellt	Integrität durch den Programmablauf sicher zu stellen
Optimiert auf transaktionale Datenkonsistenz, parallel konkurrierende Schreib- und Lesezugriffe	Optimiert auf Datendurchsatz und Performanz, asynchrone Schreib- und Lesezugriffe[23]
Für bestimmte Geschäftsprozesse und Transaktionen implementiert	Anwendungsfälle meist stark analytischer Art
Größe von Dateneinheiten eher gering (z. B. Datensatz)	Sehr große Dateneinheiten möglich
Datenabfrage und -manipulation meist durch SQL	Spezifische Abfragesprachen wie MapReduce

Tabelle 4: Vergleich der Schwerpunkte transaktionaler Systeme und Big Data

19 z. B. Internet, Globalisierung
20 z. B. OLTP
21 z. B. Data Warehouse
22 z. B. in relationalen Datenbanktabellen
23 Eventual Consistency – »Irgendwann Konsistenz« häufig ausreichend



5.2 Analytische Systeme: Data Warehouse und Business Intelligence

Die steigenden technischen Anforderungen bei der Analyse von Daten wurde frühzeitig durch neue Tools adressiert und unter den Schlagworten Data Warehouse und Business Intelligence zusammengefasst. Aufsetzend auf den bestehenden transaktionalen Systemen erlauben diese Tools einen Blick auf die Daten, der direkt in die unternehmerische Entscheidungsfindung eingehen soll. Aus ihrer Entwicklung heraus eignen sich diese Systeme vor allem für strukturierte Daten, die auf SQL-Abfragen und Tabellenstrukturen optimiert sind. Sie sind in ihrem Einsatz auf wenige, leistungsfähige und auch teure Computer optimiert.

Allerdings geht diese Blickweise meist damit einher, dass aus den Originaldaten periodisch eine verdichtete und bereinigte Kopie erstellt wird[24]. Eine feine Detaillierung der Daten wird absichtlich weg gelassen, um die Geschwindigkeit der Auswertungen zu erhöhen. Eine Anpassung der Verdichtungs-Algorithmen in Data Warehouse Tools ist in Unternehmen aufwändig, vor allem durch die zugrundeliegenden Prozessabläufe der Software-Entwicklung und Qualitätssicherung.

Die meisten Systeme dieser Art verarbeiten heute Datenmengen im niedrig-stelligen Terabyte-Bereich – eine Größenordnung, die sich auch bei kleinen Big-Data-Ansätzen findet. Auch in ihrer Aufgabenstellung überschneiden sich die existierenden analytischen Systeme und Big Data. Die Schwerpunkte sind jedoch deutlich unterschiedlich (vgl. Tabelle 5) und spiegeln sich in den genutzten technologischen und organisatorischen Ansätzen wider.

Ein punktueller Einsatz von Big-Data-Techniken kann dabei bestehende Business-Intelligence-Lösungen ergänzen und z. B. die Auswertung großer Bestandsdaten[25] beschleunigen oder die Aktualisierungshäufigkeit von Bewegungsdaten[26] drastisch erhöhen.

Analytische Systeme	Big Data
Zentrale Datenhaltung, alle Daten müssen exakt zueinander passen	Daten existieren an mehreren Stellen, Ungenauigkeiten sind akzeptabel
Qualitativ hochwertige Daten	Einfachheit der Nutzung
Strukturierte, bereinigte und aggregierte Daten	Verarbeitung der Rohdaten mit vielen unterschiedlichen Formaten
Wiederkehrende Berichte	Interaktion in Echtzeit
Periodische Erstellung	Optimiert für Flexibilität
Zentralistische Organisation	Heterogene, dezentrale Organisation

Tabelle 5: Analytische Systeme und Big Data - Vergleich der Schwerpunkte

5.3 Dokumentenmanagement

Für den Unternehmenseinsatz entstanden auch datenbankgestützte Systeme mit einer Spezialisierung auf die Verwaltung elektronischer Dokumente und freier digitaler Formate und Inhalte, d. h. hauptsächlich für unstrukturierte oder polystrukturierte Daten[27]. Zu dieser Kategorie gehören Dokumenten-Management-Systeme (DMS) und Enterprise-Content-Management (ECM)- Systeme, die den alltäglichen unternehmensinternen Workflow unterstützen. Der dokumentgestützte Workflow stellt bereits einen wesentlichen Unterschied (vgl. Tabelle 6) zwischen DMS und Big Data dar, letztere dienen der Analyse meist flüchtiger Daten und bieten keine Unterstützung z. B. für den rechtssicheren Umgang mit Dokumenten innerhalb eines Unternehmens.

[24] ETL: Extract – Transform – Load
[25] Data at Rest
[26] Data in Motion
[27] z. B. Texte, Bilder, Druckdokumente

Dokumenten-Management-Systeme	Big Data
Rechtssichere Verwaltung von Dokumenten, z. B. Versionierung	Verarbeitung flüchtiger Daten
Einbindung in Workflow-Prozesse	Analyse von Datenmengen
Strukturierte Metadaten zur Beschreibung von elektronischen Inhalten	Integration vieler verschiedener und häufig wechselnder Formate
Meist für den unternehmensinternen Einsatz und einen beschränkten Nutzerkreis konzipiert	Prinzipiell unbeschränkter Nutzerkreis vorstellbar
Zugriff auf einzelne Dokumente aus einem großen Bestand	Auswertung eines großen Datenbestands mit unterschiedlichen Datenformaten
Auswertungen und Suchen meist auf Basis von Textindizes oder Metadaten	Anpassbar an sehr unterschiedliche Formate

Tabelle 6: Dokumenten-Management-Systeme und Big Data - Vergleich der Schwerpunkte

Das Wachstum unstrukturierter Daten übersteigt das Mooresche Gesetz: Die Erstellung von Dokumenten, Ton-, Bild- und Video-Aufzeichnungen durch die allgegenwärtigen Smartphones erzeugt ein immenses Datenaufkommen. Die Herausforderung von Big Data liegt darin, diese Datenformate zugänglich und für geschäftliche Prozesse und Entscheidungen nutzbar zu machen.

Dazu gehört zuerst die »Übersetzung« von unzugänglichen Daten in »ausreichend« strukturierte Daten. So können z. B. Textbeiträge aus sozialen Netzwerken in die Produktentwicklung einfließen. Es gilt dabei, viele unterschiedliche, auch unternehmensexterne Datenquellen zu erschließen und in eine Lösung zu integrieren. Um Multimedia-Datenformate wie Audio oder Video auszuwerten, bedarf es einer automatisierten Erkennung der Inhalte und einer semantischen Analyse – etwa um das Auftreten von Produktlogos in Videoclips zu erkennen.

Ein wesentlicher Baustein vieler Big-Data-Lösungen wird sich daher mit dem Zugriff auf unstrukturierte Daten und deren Transformation beschäftigen. Daten, die bereits in Textform formuliert in natürlicher Sprache vorliegen, müssen automatisch inhaltlich erschlossen und strukturiert werden. Die meisten Tool-Hersteller bieten daher Erweiterungen zur linguistischen Datenverarbeitung an, auch im Hinblick auf die Verfügbarkeit von Open-Source-Implementierungen[28].

Mit der Integration bisher brach liegender unstrukturierter Daten steigt das zu betrachtende Datenvolumen um mehr als den Faktor 1.000 – eine lokale Datenhaltung und -analyse stößt an ihre technischen und wirtschaftlichen Grenzen. Dennoch erwartet der Internet-Nutzer eine Auswertung in Bruchteilen von Sekunden, denn Zeit ist ein wesentlicher Faktor für Big-Data-Abfragen.

Neben den Big-Data-Eigenschaften von steigenden Datenvolumina und verstärkt unstrukturierten Daten, müssen Daten zukünftig in höheren Geschwindigkeiten erfasst, integriert und analysiert werden[29]. Big Data stellt somit Anforderungen an etablierte Systeme und eröffnet komplementäre Einsatzszenarien von neuen Technologien, die auf Big Data ausgerichtet sind.

5.4 Auswirkungen auf die Software-Entwicklung

Ein wesentlicher Unterschied liegt auch in der Herangehensweise an die Problemlösung und Softwareentwicklung (vgl. Abbildung 10).
Bisher steuern die Geschäftsanforderungen das Lösungsdesign. Dazu definiert der Fachbereich seine Anforderungen, legt also die Fragen fest, die mit Hilfe der Analyse beantwortet werden sollen. Die IT-Abteilung entwirft

[28] z. B. UIMA bei der Apache Foundation oder SMILA als Eclipse-Projekt

[29] z.B. Sensordaten, Mobile Apps, Twitter Tweets, Event Streams

anhand der Vorgaben eine Lösung mit einer festen Struktur und der geforderten Funktionalität. Das Ergebnis mit seinen definierten Queries und Reports wird von der Fachabteilung wieder und wieder gegen den Datenbestand in den transaktionalen und analytischen Systemen gefahren. Ergeben sich neue Anforderungen, so ist ein Redesign und Rebuild dieser Lösung durch die IT-Abteilung notwendig.

Mehr Daten liefert bessere Ergebnisse als intelligentere Algorithmen auf bestehenden Daten.

»Mehr Daten« liefert bessere Ergebnisse als intelligentere Algorithmen auf bestehenden Daten. So lässt sich heute der Big-Data-Ansatz für das Software Engineering zusammen fassen. Da im Big-Data-Umfeld der Wert der Daten häufig im Vorfeld nicht bekannt ist, zielt die Software-Entwicklung darauf, mit einem offenen Ausgang möglichst viele Daten zu untersuchen – oder besser zu erforschen:

Fachabteilung und IT-Abteilung identifizieren gemeinsam verfügbare und möglicherweise interessante Datenquellen. Die IT-Abteilung stellt dann eine flexible Plattform zur Verfügung, welche es erlaubt, diese Datenquellen zu erkunden und auszuwerten. Werden im Laufe der Analyse zusätzliche interessante Datenquellen identifiziert, ist deren Einbeziehung in den weiteren Analyseablauf auch nachträglich möglich und wünschenswert.
Anwender aus den Fachabteilungen können die Daten untersuchen, verschiedene Hypothesen testen und einen iterativen Ansatz zur Problemlösung nutzen (explorative Analyse). Für Geschäftsentscheidungen lassen sich die erzielten Ergebnisse dann mit den Ergebnissen der traditionellen Analysen kombinieren, um so zu einer insgesamt besser fundierten Entscheidung zu finden.
Big Data bedeutet daher auch, moderne Prozessansätze der Software-Entwicklung und -Projektsteuerung einzusetzen. Für Unternehmen gilt es, die Herausforderungen der neuen Prozesse, Technologien und der dazu notwendigen Qualifikationen zu meistern, ohne die bewährten Systeme und Prozesse zu gefährden oder vollständig neu zu entwickeln.

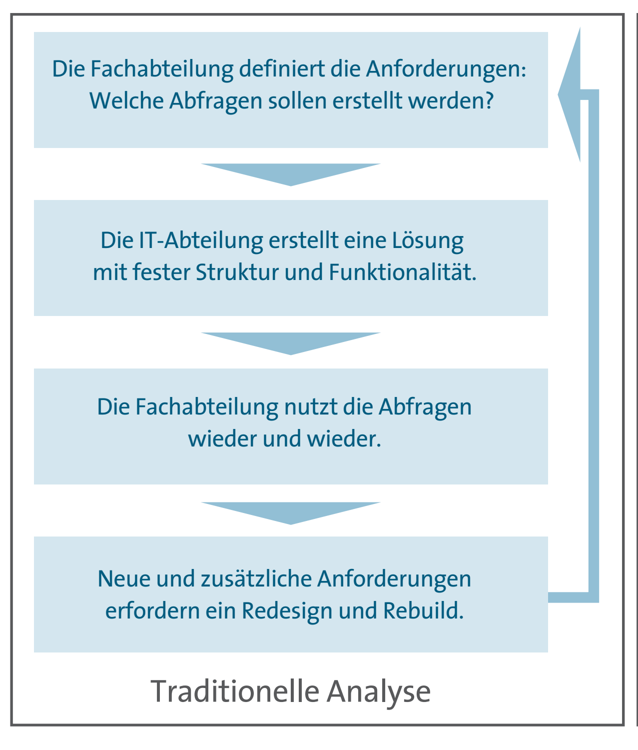


Abbildung 10: Analyseansätze für traditionelle Systeme und Big Data Systeme



5.5 Auswirkungen auf die Anwendungs-Architektur

In den letzten Jahren haben sich die proprietären Big-Data-Systeme etwas geöffnet, zuerst über die viel beachtete Veröffentlichung der Google, Inc. zu ihrem »MapReduce«-Framework[30], das ein allgemeingültiges Programmiermodell für die hochgradig verteilte Datenverarbeitung bietet.

Aus der Firma Yahoo ging 2006 das Open Source Framework »Hadoop« als eine freie Implementierung des MapReduce-Frameworks hervor. Diese Entwicklung beeinflusst bis heute Big Data maßgeblich.

Inzwischen unter der Steuerung der Apache Foundation, kann Hadoop zusammen mit neuen Open-Source-Komponenten sehr große unstrukturierte Daten (> 1000 Terabyte) auf losen Clustern günstiger Server speichern und analysieren. Mehrere Anbieter ermöglichen sogar den Einsatz innerhalb ihrer Cloud-Lösung, so dass die technische und finanzielle Hürde für eine Big-Data-Anwendung niedrig liegt und dennoch die Lösung auf große Spitzenlasten skalierbar bleibt.

Mit der zum Teil quelloffenen Verfügbarkeit von Big-Data-Lösungen stehen die vorhandenen konventionellen IT-Anwendungen unter einem immensen Druck, dieselben Herausforderungen ebenso kostengünstig und agil zu meistern – obwohl diese Anwendungen ursprünglich mit anderen Schwerpunkten entwickelt wurden (vgl. Tabelle 7). Die Hersteller und Anwender reagieren daher seit einigen Jahren mit dem verstärkten Einsatz neuer Technologien[31], die zum Teil vorhandene Anwendungen ergänzen und erweitern oder sogar vollständig durch neue Systeme ersetzen.

Transaktionale und analytische Systeme	Big Data
Struktur der Daten	Integration verschiedenster Daten
Konzentration auf hochwertige Daten	Umgang mit variablen Daten und Inhalten
Identifikation wiederholbarer Operationen und Prozesse	Schnelle Anpassbarkeit bis hin zu Einzelfall-Analysen
Definierte Anforderungen der Fachanwender	Iterative und explorative Analyse als Reaktion auf fehlende oder wechselnde Anforderungen
Optimiert für schnellen Zugriff und bestehende Analysen	Optimiert für Flexibilität

Tabelle 7: Schwerpunkte bei der Anwendungs-Analyse

Einhergehend mit der Open-Source-Bewegung für die verteilte Datenverarbeitung entwickeln sich Lösungen, um Daten ohne feste Struktur zu speichern. In Anspielung auf die seit langem etablierten starren relationalen Datenbanken mit ihrer Abfragesprache SQL hat sich der Überbegriff »NoSQL« gebildet – »not only SQL«. Die Abfragesprache SQL wird jedoch auch in diesen Systemen weiterhin relevant sein und von den Anwendern genutzt. Die einzelnen Anbieter von Speicherlösungen unterscheiden sich dabei z. B. in der eingesetzten Speicherstruktur: Ist die Geschwindigkeit einer Anwendung das Hauptkriterium[32], dann bieten die meisten Hersteller als Ergänzung In-Memory-Technologien an. So werden bestehende Anwendungen drastisch schneller und neue Anwendungsfelder wie Big Data können u. U. mit bestehenden Anwendungen realisiert werden. Besonders die analytischen Systeme[33] profitieren durch den Geschwindigkeitszuwachs und erlauben z. B. die interaktive visuelle Arbeit mit den Geschäftsdaten für Finanz- oder Vertriebsanalysen.

30 Vgl.: http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/de//archive/mapreduce-osdi04.pdf (Abruf: 26.07.2012)

31 z. B. Datenbank-Appliance, Hadoop-Integration

32 bei überschaubaren Datenmengen oder bei der Arbeit auf verdichteten Daten

33 Data Warehouse und BI



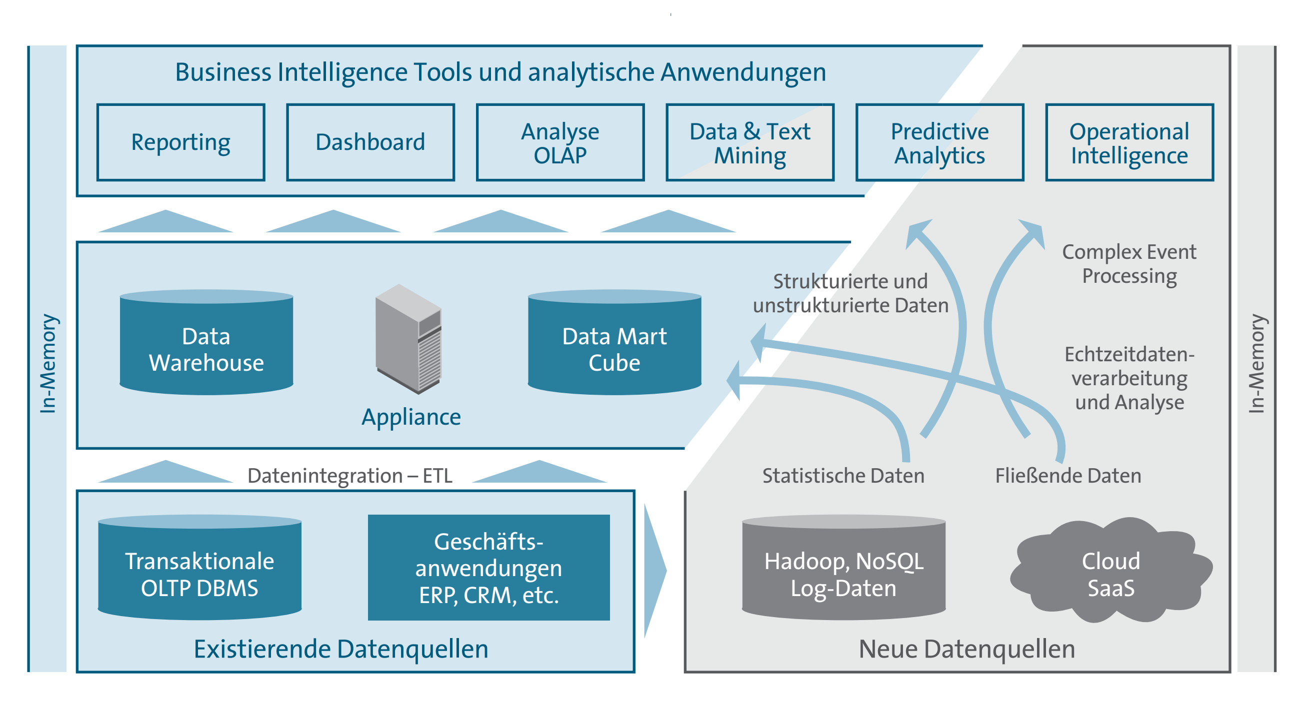


Abbildung 11: Integrierte Anwendungslandschaft mit traditionellen Systemen und Big-Data-Lösungen

Da Daten zukünftig mehr und mehr nicht nur persistent in einem Unternehmen liegen, sondern dynamisch außerhalb erzeugt werden und riesige Datenmengen in geringen Zeitabständen in das Unternehmen »hineinströmen«[34], werden neue technische Lösungen wie das Complex Event Processing (CEP) verstärkt im Kontext von Big Data eingesetzt.

In Unternehmen wird zukünftig eine Kombination von konventionellen und neuen Technologien zum Einsatz kommen (vgl. Abbildung 11), um Big-Data-Lösungen bereitzustellen. Beispielsweise können Hadoop-Cluster dazu verwendet werden, sehr große Datenmengen für ein Data Warehouse bereitzustellen; bestehende Datenbanksysteme lassen sich durch den Einsatz von In-Memory Technologien beschleunigen. Ziel für jedes Unternehmen sollte es sein, die Technologien zur Adressierung von Big-Data-Anforderungen so einzusetzen, dass ein betriebswirtschaftlicher Mehrwert entsteht und das Risiko für die bestehenden Information-Management-Systeme möglichst gering gehalten wird.

5.6 Big Data als Fortentwicklung vorhandener Technologien

Die beschriebenen und bereits in den meisten Unternehmen vorhandenen Technologien erlauben jeweils einen graduellen Übergang auf eine Big-Data-Lösung oder eine Integration mit Big-Data-Techniken. Ausgehend von der bestehenden Anwendungslandschaft lassen sich damit erste Schritte in Richtung Big Data gut ableiten (vgl. Tabelle 8).

[34] z. B. Sensordaten, Twitter Tweets

Aspekt	Schritt in Richtung Big Data: Prüfung des Einsatzes von...
Datenintegration	... linguistischer Datenverarbeitung zur Auswertung von Textdokumenten.
Verarbeitungsgeschwindigkeit und Skalierbarkeit	... In-Memory-Technologien oder dedizierten Appliances für transaktionale Systeme
Analyse und Speicherung großer Datenmengen	... Hadoop-Systemen zusammen mit bestehenden Data-Warehouse-, BI- und ETL-Systemen
Entscheidungsfindung	... CEP-Technologien zur Verarbeitung bestehender Datenströme
Investitionskosten	... standardisierter, eventuell quell-offener Software und von bestehenden, direkt abrufbaren Cloud-Lösungen
Entwicklungs- und Analysezyklen	... explorativen Analyseansätzen und agilen Projektmanagement-Methoden für die Weiterentwicklung bestehender Systeme

Tabelle 8: Schritte in Richtung Big Data

5.7 Auswirkungen von Big Data auf benachbarte Disziplinen

Betrachtet man die langfristigen Trends der Informationstechnik, so könnte Big Data das erste Anzeichen eines weiteren fundamentalen Umbruchs darstellen. Lag am Anfang der Entwicklung der Schwerpunkt in der IT auf Hardware, so ist seit langem die Software der dominierende Faktor. Dies geht so weit, dass heute Softwaresysteme einen wesentlichen Wertbeitrag in allen Unternehmen liefern und als solche durchaus auch in die Bilanz des jeweiligen Unternehmens eingehen. Das Aufkommen der großen weltweiten Internet-Unternehmen zeigt jedoch, dass die Daten selbst der entscheidende Erfolgsfaktor werden.

Für jedes Unternehmen stellt sich daher die Frage, ob nicht bereits in naher Zukunft der Umgang mit den Daten wichtiger wird als die betriebene Software-Anwendung. Um die Bedeutung und Auswirkung von Big Data auf den eigenen Geschäftserfolg zu verstehen, gilt es zunächst, den Stellenwert der Daten zu bestimmen:

- Welche Art von neuen Daten[35] haben einen wichtigen Einfluss auf den zukünftigen Unternehmenserfolg und die Geschäftsprozesse?
- Liegen diese Daten im Unternehmen vor oder werden sie extern zur Verfügung gestellt?
- In welchen Datenstrukturen liegen die Daten vor und wie[36] müssen diese neuen Daten verarbeitet werden, um den maximalen Geschäftsnutzen zu erzeugen?
- Welchen Einfluss haben Big-Data-Ausprägungen auf den Daten-Lebenszyklus[37]?
- Sind heute IT-Systeme im Einsatz, die mit den verschiedenen Merkmalen von Big Data umgehen können? Wo sind Lücken erkennbar?
- Haben die Mitarbeiter das notwendige Wissen, das Big-Data-Phänomen für das Unternehmen und für seine IT-Systeme zu bewerten?

Auf diese Weise beginnt die Weiterentwicklung der internen IT mit einer Bestimmung des Daten-Portfolios. Erst die angestrebte Datennutzung ergibt die Software-Architektur, die in meisten Unternehmen die neuen Technologien und Ziele mit einem gut funktionierenden, vorhandenen Anwendungspark integrieren muss.

35 z. B. Social Media, Sensordaten, mobile Endgeräte und Apps
36 z. B. Echtzeitdatenauswertung, semantische Analyse
37 z. B. erfassen, integrieren, modellieren, speichern, auswerten, visualisieren, archivieren

6 Nutzung unstrukturierter Daten in Big-Data-Verfahren

Die automatisierte inhaltliche Erschließung auf Basis von semantischen und multimedialen Analyseverfahren erweitert den Blick einer Organisation auf relevante Datenbestände. Sie überführt unstrukturierte Daten automatisiert in eine strukturierte Form. Die Analyseverfahren heben die Datenqualität auch für unstrukturierte Daten auf ein Niveau, das weitere quantitative und qualitative Analysen ermöglicht. Mit dieser Struktur können durch iterativ verfeinerte Abfragen genau die Daten ermittelt werden, welche für Entscheider relevant sind. Entscheidungen werden so fundiert getroffen, und Unternehmen erarbeiten sich einen Wettbewerbsvorteil.

Eine wesentliche, historisch bedingte Herkunft von Big-Data-Anwendungen sind klassische Business-Intelligence-Anwendungen. Diese fokussieren auf strukturierte Daten, wie sie beispielsweise in ERP-Systemen und Datenbanken vorkommen. Quantitative Auswertungen – in grafischen und tabellarischen Reports zusammengefasst – stehen dabei im Vordergrund, sind heute längst fester Bestandteil der IT-Strategie in Unternehmen und allgemein akzeptiert.

In Big-Data-Anwendungen werden durch die Hinzunahme weiterer interner und externer Datenquellen Prozesse beschleunigt und Entscheidungen auf eine besser fundierte Basis gestellt. Allerdings liegen diese Datenquellen oftmals in unstrukturierter Form[38] vor.

Es wird eingeschätzt, dass für ein Unternehmen 1.000 Mal mehr unstrukturierte als strukturierte Daten relevant sind[39]. Bei einer 2011 durchgeführten Umfrage[40] gaben etwa 50% der Beteiligten an, dass unstrukturierte Daten 40% oder mehr ihres Datenaufkommens ausmachen. Beide Angaben zeigen, dass unstrukturierte Daten einen wesentlichen Anteil am Datenaufkommen haben. Durch eine inhaltliche Erschließung werden diese unstrukturierten Daten ebenfalls quantitativ auswertbar und ermöglichen so verlässliche Analyse- und Planungsszenarien.

In der Studie der Computerwoche gaben nur ca. 4% der Beteiligten an, dass ihr Unternehmen voll und ganz über die richtigen Werkzeuge verfügt, um mit unstrukturierten Daten umzugehen und diese in verschiedenen Szenarien sinnvoll zu nutzen.

Competitive Intelligence

Ein Anwendungsbeispiel zur Analyse unstrukturierter Daten findet sich bei der Competitive Intelligence[41] aus öffentlich zugänglichen Inhalten. Die so erzielten Erkenntnisse über Mitbewerber und Technologien bilden einen wesentlichen Wettbewerbsvorteil und ergänzen allgemeine Studien – beispielsweise von Marktforschungsunternehmen. Die dazu nötigen Informationen sind bereits im Internet verfügbar[42].

---

38 beispielsweise in Texten, Geschäfts- und PDF-Dokumenten, Call-Center-Notizen, Sensor- und RFID-Daten, Webinhalten, Intranet, Social Media, Video, Audio und Grafiken

39 http://www.forrester.com/Understanding+The+Business+Intelligence+Growth+Opportunity/fulltext/-/E-RES59995

40 Vgl.: Big Data. Eine Marktanalyse der Computerwoche, München, November 2011

41 Vgl. Einsatzbeispiel im Abschnitt 10.2.1

42 Ein Konsortium unter der Leitung der TU Berlin entwickelt im Forschungsprojekt MIA (Marktplatz für Informationen und Analysen) einen Marktplatz für die Verwertung des Datenbestands des deutschsprachigen Webs. Die entwickelten Dienste für die gemeinschaftliche Speicherung, Analyse und Verwertung der Daten ermöglichen völlig neuartige Geschäftsmodelle mit Informationen und Analysen als elektronisch handelbarem Gut.
Vgl.: http://www.mia-marktplatz.de/

Einige Beispiele hierfür sind:

- Fachzeitschriften für Produkterfahrungen, Kundenbedürfnisse und Marktanalysen
- Soziale Medien, um Feedback aus der Anwendung zu ermitteln
- Webseiten, Pressemitteilungen und Geschäftsberichte von Mitbewerbern für eine Mitbewerberanalyse
- Wissenschaftliche Veröffentlichungen, um neue Grundlagentechnologien zu bewerten
- Patentanmeldungen, um das Engagement von Mitbewerbern und technische Umsetzungsmöglichkeiten zu ermitteln
- Politische Entscheidungen, Gesetze und Rechtsprechung zur Ermittlung der Rahmenbedingungen.

Ohne technische Unterstützung werden diese Analysen aufgrund des damit verbundenen Aufwandes nur sporadisch und auf das jeweilige Themengebiet fokussiert durchgeführt. So gehen wesentliche Informationen verloren: Der Einfluss einer neuen gesetzlichen Regelung auf laufende Entwicklungsprojekte bleibt unerkannt, oder neue Produkte von Mitbewerbern werden nur mit zeitlicher Verzögerung erkannt. Feedback aus der Anwendung fließt ebenfalls mit zeitlicher Verzögerung in den Entwicklungsprozess ein.

Das Wissen um die relevanten Quellen und zur Erkennung relevanter Inhalte ist in der Organisation vorhanden. Durch Verfahren zur inhaltlichen Erschließung wird die in einem Unternehmen vorhandene Domänenkompetenz weiter operationalisiert und auf ein neues Niveau gehoben:

- Die Menge der analysierten Informationen wird – bei verbesserten Ergebnissen – drastisch erhöht. Durch die umfassende Sicht auf relevante Quellen werden die Ergebnisse belastbarer.

- Die Analysen können periodisch wiederholt und so Trendaussagen abgeleitet werden.

Kombination aus Linguistik und Semantik

Um die Dokumente – und die darin enthaltenen Freitexte – richtig und zieladäquat zu analysieren, ist eine Kombination aus Linguistik und Semantik nötig. Die linguistischen Verfahren analysieren die Zusammenhänge in den Texten – beispielsweise »Firma A entwickelt Technologie X« oder die Umsatzzahlen eines Unternehmens. Durch die Linguistik wird dieser Zusammenhang mit hoher Sicherheit auch bei komplexeren Satzstrukturen wie Nebensätzen, Negierungen oder Passiv-Formulierungen erkannt. Eine einfache, schlüsselwortbasierte Analyse (keyword-search), ob die jeweiligen Begriffe in einem Dokument zusammen vorkommen, führt gerade bei langen Dokumenten wie Geschäftsberichten zu Treffern, die keinen inhaltlichen Zusammenhang ausdrücken. Auch kann ein Teil der linguistischen Analyse »offen« formuliert werden, um so bisher unbekannte Inhalte zu erkennen.

Wissensmodelle der Semantik

Ein Beispiel ist die Regel <Firma><entwickelt><X>, wobei X ein Kandidat für eine mögliche neue Technologie ist. Durch die Kombination von Linguistik und Semantik wird die Information weiter sinnvoll verdichtet. In den Wissensmodellen der Semantik sind die für die Analyse relevanten Begriffe und deren Zusammenhänge beschrieben – wie Produktbezeichnungen oder Namen von Mitbewerbern.

Bei den Mitbewerbern können so Patentanmeldungen durch unterschiedliche Organisationseinheiten oder Umbenennungen des Firmennamens berücksichtigt werden. Technologien werden zu inhaltlichen Gruppen oder nach ihren Einsatzzwecken zusammengefasst, so dass auch Auswertungen über Technologiecluster möglich sind. Die Inhalte der Wissensmodelle sind innerhalb der Firma bereits vorhanden – nicht nur in den Köpfen der Mitarbeiter, sondern auch in verschiedenen Systemen wie der Produktdatenhaltung oder in eCommerce-Systemen. Durch Importprozesse können daher diese Informationen in das Wissensmodell überführt werden, um so die Analyse zu

unterstützen. Weitere frei verfügbare Wissensmodelle aus dem Open-Linked-Data-Bereich – wie zum Beispiel Geonames[43]– bilden eine andere Quelle für die inhaltliche Analyse auf Basis von Wissensmodellen. Webbasierte Werkzeuge unterstützen zusätzlich bei der verteilten Erfassung des Wissensmodells über Standortgrenzen hinweg. Linguistik und Semantik zusammen strukturieren damit die vorliegenden Freitexte so, dass quantitative Auswertungen durchgeführt werden können – analog zu klassischen OLAP-Anwendungen. So können Dashboards definiert werden, um neue Technologien oder Märkte zu bewerten. Die im Wissensmodell definierten Zusammenhänge ermöglichen die Analyse nach verschiedenen Kriterien – wie Quellen, Branchen oder Mitbewerber. Wenn nötig, kann so schnell und interaktiv die Datenbasis eingeschränkt werden. Der Benutzer erhält eine auf seine Bedürfnisse abgestimmte Übersicht und muss nicht mehr zwingend einzelne Beiträge detailliert analysieren. Weiterhin beschleunigen die durch die Übersicht gewonnenen neuen Erkenntnisse bei Bedarf die Analyse von Einzelbeiträgen.

Die so erzielte inhaltliche Verdichtung erlaubt damit zielgerichtete, auf konkrete Fragestellungen angepasste Analysen bei geringeren Kosten. Sie schafft damit einen substantiellen Wettbewerbsvorteil.

Analyse von Bildern, Videos und Audiodateien

Die Analyse von unstrukturierten Daten ist dabei nicht zwingend auf textbasierte Daten beschränkt. Wenn Sprachaufzeichnungen automatisiert in Text überführt werden, können zum Beispiel auch Podcasts oder die Tonspur von Videobeiträgen in die Analyse einbezogen werden. Die Erkennung von Logos in benutzergenerierten Videos auf Videoplattformen erlaubt eine Bestimmung der Reichweite von Marketingkampagnen. Die Erkennung von chemischen Strukturformeln in Patentdokumenten erlaubt es, Trends in den Entwicklungsaktivitäten von Marktteilnehmern oder ähnlich gelagerte eigene Entwicklungen zu identifizieren.

Ergebnisse des THESEUS-Programmes

Die zuvor genannten Analyseverfahren sind alles Beispiele von Technologien, welche im Rahmen von THESEUS[44] entwickelt oder zur praktischen Einsatzreife gebracht wurden. Neben dem Bundeswirtschaftsministerium waren rund 60 Partner mit dem Ziel beteiligt, den Zugang zu Informationen zu vereinfachen, Daten zu neuem Wissen zu verknüpfen und neue Dienstleistungen im Internet zu ermöglichen.[45]

Hierbei konnte THESEUS auf die jahrzehntelange Exzellenz von Wissenschaft und Anwendung im Bereich der Semantik aufbauen. Linguistische Verfahren für Deutsch und andere Sprachen in unterschiedlichen Szenarien konnten weiter ausgebaut und bis zur praktischen Anwendung weiterentwickelt werden. Neue Verfahren zur Analyse von Bildern, Videos und Audiodateien erlauben es, weitere Medienformate in die inhaltliche Erschließung einzubeziehen. Mit THESEUS sind wesentliche Beiträge zur Transformation des Wirtschaftsstandortes Deutschland in eine Wissensgesellschaft geschaffen worden. Innovative Technologien für das Internet der Dienste wurden durch die Förderung von THESEUS – mit rund 50 Patenten und 130 laufenden Systemen – möglich.

---

43 Geonames.org

44 THESEUS ist das bisher größte IT-Forschungsprojektes Deutschlands. Vgl.: www.theseus-programm.de

45 Das Competitive-Intelligence-Beispiel (vgl. 10.2.1) nutzt neben den Verfahren zur inhaltlichen Erschließung auch Technologien zur Hochskalierung auf unstrukturierten Daten.

THESEUS-Beiträge für Big-Data-Szenarien

Auch für die verschiedenen Herausforderungen bei der inhaltlichen Erschließung in Big-Data-Szenarien wurden in THESEUS industriell einsatzfähige Lösungen entwickelt.

- Verschiedene Datenquellen mit hohem Datenvolumen (Volume) müssen performant angebunden werden (Velocity). Hier werden Basiskonnektoren – zum Beispiel für Web, Datei oder Datenbanken – benötigt, um Anbindungen an relevante Quellen schnell und effizient durchzuführen. Für den Import umfangreicher Datenmengen muss der Import innerhalb der zur Verfügung stehenden technischen Infrastruktur skalieren.

- Je nach Art der Quelle muss eine Konvertierung für die weitere Verarbeitung stattfinden – beispielsweise die Extraktion der textuellen Inhalte aus einem Office-Dokument (Variety). Die Unterstützung von Standard-Formaten ist hier ein Kriterium für eine effiziente Anbindung einer Datenquelle.

- Die darauf aufbauende inhaltliche Erschließung findet durch eine oder mehrere spezialisierte Komponenten statt – in Abhängigkeit davon, welche Inhalte für die Analyse relevant sind. Die jeweiligen Komponenten müssen daher in frei definierbaren Arbeitsabläufen miteinander kombiniert werden. Bei der Definition ist auf Flexibilität zu achten, da sich die Anforderungen an die Arbeitsabläufe ändern – und mit neuen Datenquellen neue, teilweise ähnliche Arbeitsabläufe benötigt werden.

Content-Verarbeitungs-Frameworks

Hierzu wurde mit »SMILA[46] – Unified Information Access Architecture«, als Projekt der Eclipse Foundation, ein Open Source Framework entwickelt, um die verschiedenen Verfahren hochskalierbar und flexibel miteinander zu integrieren. SMILA vereint in einem Framework die effiziente Verarbeitung umfangreicher Datenbestände durch asynchrone Workflows und die schnelle Abfrage durch synchrone Verarbeitung in Pipelines.

Unabhängig von der Auswahl des eingesetzten Frameworks – mit UIMA[47] als einer weiteren Alternative – ist darauf zu achten, dass auch die eingesetzten Verfahren an sich skalierbar sind und sich auf die vorhandenen Rechnerknoten verteilen. Idealerweise findet zwischen den einzelnen Instanzen des Analyseverfahrens nach einer Initialisierung keine weitere Kommunikation mehr statt. Neben der Skalierbarkeit ist ein wichtiges Entscheidungskriterium die Performanz des Analyseverfahrens – da die jeweiligen Verfahren in Big-Data-Anwendungen auf eine umfangreiche Datenmenge angewendet werden, können sich Investitionen in die Performanz von Verfahren durch geringeren Ressourceneinsatz bezahlt machen.

Die Flexibilität derartiger Content-Verarbeitungs-Frameworks in Verbindung mit einer hochskalierenden Infrastruktur ermöglicht es den einsetzenden Unternehmen, Applikationen schnell und effizient an neue Anforderungen – wie die Erschließung einer neuen Datenquelle oder neue Analyseverfahren – anzupassen.

46 http://www.eclipse.org/smila/

47 http://uima.apache.org/

7 Big Data – Praxiseinsatz und wirtschaftlicher Nutzen

Kapital, Arbeitskraft und Rohstoffe gelten als die klassischen Produktionsfaktoren der Wirtschaft. In der digitalen Welt treten Daten jeglicher Ausprägung als vierter Produktionsfaktor[48] hinzu.

Big Data gewinnt zunehmend an Bedeutung, weil das Volumen der zur Verfügung stehenden Daten wie auch die Zahl der Datentypen wächst. Mit neuen Hard- und Software-basierten Verfahren lässt sich die Flut der meist unstrukturierten Daten in einen sinnvoll nutzbaren Produktionsfaktor verwandeln. Big-Data-Analysen generieren erheblichen Mehrwert und beeinflussen maßgeblich die Strukturen von Organisationen und Unternehmen sowie das Management.

Der wirtschaftliche Nutzen von Big Data lässt sich in einigen Funktionsbereichen besonders eindrucksvoll belegen. Hierzu gehören insbesondere Marketing und Vertrieb, Forschung und Entwicklung, Produktion, Service und Support, Distribution und Logistik, Finanz- und Risiko-Controlling sowie Administration und Organisation.[49]

- Big Data erleichtert es Marketing- und Vertriebsabteilungen, Produkt- und Service-Angebote zunehmend auf Kundensegmente oder einzelne Kunden zuzuschneiden und Streuverluste im Marketing zu vermindern.
- Ein hohes Potenzial für den Einsatz von Big Data schlummert in der Wissenschaft sowie in der betrieblichen Forschung und Entwicklung. Meteorologie, Klimaforschung, Lagerstätten-Erkundung von Rohstoffen, Atomphysik und die Vorhersage von Epidemien profitieren gleichermaßen von Fortschritten im Bereich Big Data. In der Entwicklung der nächsten Produktgeneration helfen Social-Media-Analysen und die Auswertung von Sensordaten der zurzeit im Einsatz befindlichen Produkte.
- Mit dem Internet der Dinge oder M2M-Kommunikation können produzierende Unternehmen ihre Fertigungs-, Service- und Supportprozesse optimieren. Dafür erfassen Sensoren an Produkten und entlang von Produktions- und Lieferketten Daten - auch im späteren Betrieb. Viele Unternehmen arbeiten daran, die verschiedenen Unternehmensbereiche zu verknüpfen und in die Optimierung auch Zulieferer und Partner einzubinden.
- In Distribution und Logistik geht es um nachhaltige Kostensenkung auf dem Wege einer stärkeren Vernetzung von Fahrzeugen mit der Außenwelt. Immer mehr Fahrzeuge werden mit Sensoren und Steuerungsmodulen ausgestattet, die Fahrzeugdaten wie den Benzinverbrauch, den Zustand von Verschleißteilen oder Positionsdaten erfassen und in Datenbanken übertragen. Mit diesen Daten können Disponenten zeitnah Transporte planen, gegebenenfalls Routen und Beladung ändern, Wartungskosten und Stillstandzeiten minimieren.
- Das Finanz- und Risiko-Controlling profitiert u.a. von neuen Möglichkeiten im Bereich Betrugserkennung und Risikomanagement. Bei der Betrugserkennung steht in erster Linie eine möglichst vollständige Sammlung und Beobachtung relevanter Handlungen im Vordergrund. Das Risikomanagement wird durch hochkomplexe Berechnungen unterstützt.

---

[48] Mitunter werden Daten lediglich als neuer Rohstoff gesehen, der erst durch Analyse veredelt wird und dann als Baustein für Entwicklungs-, Produktions- und Vermarktungsprozesse und grundsätzliche Geschäftsentscheidungen bereitsteht.

[49] diese Funktionsbereiche werden im Kapitel 10 des Leitfadens Einsatzbeispiele vorgestellt.

7.1 Marketing & Vertrieb

Zu den wichtigsten Einsatzgebieten von Big Data in den Bereichen Marketing und Vertrieb gehören:

- Kostenreduzierung in Marketing und Vertrieb
- Erhöhung des Umsatzes bei Verkaufsvorgängen
- Markt- und Wettbewerberanalysen
- Erhöhung von Point of Sales Umsätzen
- Management von Kundenabwanderungen.

Kostenreduzierung im Marketing und Vertrieb

Big Data erleichtert es Marketing- und Vertriebsabteilungen, Produkt- und Service-Angebote zunehmend auf Kundensegmente oder einzelne Kunden zuzuschneiden und Streuverluste im Marketing zu vermindern.

In diesem Szenario wird der Erfolg von Maßnahmen und Online-Kampagnen gemessen - also die Frage beantwortet, wie hoch die zusätzlichen Umsätze durch bestimmte Maßnahmen sind. Zu diesem Zweck wird eine sehr große Zahl von Daten zu Nutzerverhalten im Netz erhoben und ausgewertet.

Erhöhung des Umsatzes bei Verkaufsvorgängen

Dem Handel eröffnen sich Cross-Selling-Potenziale, indem Einzelhändler typische Muster für Kaufentscheidungen identifizieren. Online-Händler erhöhen mit solchen Analysen den Umsatz pro Kaufvorgang.[50] Cross Selling könnte aber auch über einen Kunden bekannte Daten wie Transaktionen oder aktuelle Standortdaten verwenden und mit weiteren – beispielsweise demographischen – Daten in Echtzeit in Beziehung setzen. Händler wären in der Lage, einem Kunden zu einem bestimmten Zeitpunkt an einem Ort spezifische Angebote zu unterbreiten.[51]

Markt- und Wettbewerbsbeobachtung

Auch die Markt- und Wettbewerbsbeobachtung lässt sich mit Big-Data-Analysen deutlich erweitern. So können Informationen von den Internetseiten der Wettbewerber, aus der Fach-, Wirtschafts- und Lokalpresse oder von Fachportalen in die Auswertung einfließen. Dazu kommen Social-Media-Inhalte aus Facebook, Blogs, internen Wikis oder Foren.

Solche Daten werden mit verschiedenen intelligenten Verfahren erhoben. Mit Screen Scraping lassen sich Texte aus Computerbildschirmen auslesen. Auch die semantische Auszeichnung von Webinhalten gibt Hinweise auf den Stellenwert einer Information. Groß oder fett ausgezeichnete Webinhalte sind in der Regel höher zu bewerten als Fließtext. Die Ergebnisse von Suchmaschinen lassen sich ebenfalls auswerten. Die Herkunft der Treffer gibt Auskunft über regionale Märkte und Segmente und liefert Hinweise zur Qualität der Informationen von Wettbewerbern sowie die Güte von Quellen. Aus all diesen strukturierten und unstrukturierten Daten entstehen Reports über Märkte und Wettbewerber, die umfassender und aktueller sind als konventionell erstellte Berichte. Big-Data-Analysen verbessern insgesamt die Basis für die Entwicklung fundierter Unternehmens-, Produkt- und Marktstrategien.

50 Wer etwa bei Amazon ein Buch bestellen will, bekommt weitere Produkte angezeigt, für die sich andere Käufer zuvor entschieden haben. Ein Käufer des Star-Wars-DVD-Pakets könnte sich beispielsweise für ein passendes Buch zur Filmproduktion, für ein Laserschwert oder für andere Science-Fiction-Filme interessieren.

51 Der Fan der Krimibuchreihe um Commissario Montalbano erhält dann zum Beispiel in einem Restaurant an seinem aktuellen Aufenthaltsort einen Rabatt für ein spezielles Montalbano-Menü.

Erhöhung von Point-of-Sales-Umsätzen

Point-of-Sales-basiertes Marketing – auch Location-based Marketing genannt – wird für die individuelle In-Store-Ansprache von Kunden genutzt. Wer an der Kasse seinen Einkauf bezahlt, erhält dann Informationen und Rabatthinweise für Artikel, die den Einkauf ergänzen könnten.
Auf Basis aller Kundendaten lassen sich durch Verknüpfungen Verhaltens- und Kaufmuster aus einem scheinbaren Datenchaos identifizieren, die ohne Big-Data-Analysen kaum aufgefallen wären.
Die Mikrosegmentierung bildet die Voraussetzung, dass Marketingabteilungen ihre Werbemaßnahmen letztlich individuell, also mit deutlich weniger Streuverlusten als bisher adressieren.

Management von Kundenabwanderungen

Big Data kann auch genutzt werden, um frühzeitig drohende Kundenabwanderungen zu identifizieren und entgegen zu steuern. So können z. B. Mobilfunkanbieter im Prepaid-Geschäft auswerten, bei welchen Kunden es in der Vergangenheit zu Netz- oder Qualitätsproblemen gekommen ist und bei welche Kunden aufgrund ihres Telefonverhaltens eine Abwanderung droht. Hier kann gezielt durch Rabattpakete oder Verkaufsmaßnahmen entgegen gewirkt werden.

Finanzinstitute verbessern ihr Zielgruppenmarketing

Eine anderes Beispiel sind Finanzinstitute, die ihr Zielgruppenmarketing verbessern wollen. Dafür ziehen sie Informationen wie Kunden-, Konten- oder Antragsdaten aus unterschiedlichen Systemen in einer Datenbank zusammen. So lässt sich etwa feststellen, welcher Depotkunde wie viele Transaktionen tätigt. Wer nur selten aktiv mit Anlagepapieren handelt, dem muss die Bank dann kein Angebot unterbreiten, das auf Vielhändler zugeschnitten ist. Auch bei der Neukundenakquise kann eine Datenanalyse helfen, Kampagnen besser auf einzelne Empfänger abzustimmen. Dafür werden unter anderem Rücklaufquoten pro Mailing und Kunde ermittelt und daraufhin Verteilerlisten angepasst. Bei einer großen Zahl von Mailings pro Jahr werden auf diese Weise die Kosten erheblich verringert. Es wird verhindert, Verbrauchern per Post Privatkredite anzubieten, wenn diese noch nie bereit waren, für einen Einkauf Schulden zu machen.

Beispiel Einzelhandel

Im Einzelhandel lässt sich das Potenzial von Big Data besonders überzeugend illustrieren: Cross-Selling, Location-based Marketing, In-Store-Verhaltensanalyse, Mikrosegmentierung von Kunden, Sentiment-Analyse und schließlich ein verbessertes Kundenerlebnis quer durch möglichst alle Marketing- und Vertriebskanäle sind fünf Bereiche, bei denen die Auswertung umfangreicher Datenbestände von kritischer Bedeutung ist. Cross-Selling etwa verwendet alle über einen Kunden bekannten Daten - Demographie, Transaktionen, Präferenzen und bekannte Standorte in Echtzeit, um den Umsatz pro Kaufvorgang zu erhöhen.

7.2 Forschung und Produktentwicklung

Unter den relevanten Einsatzgebieten von Big Data in Forschung und Produktentwicklung sind zu nennen:

- Echtzeit-Auswertung von komplexen Daten aus wissenschaftlichen Experimenten
- Produktneuentwicklungen und -verbesserungen
- Social-Media-Trendanalysen für neue Produktideen
- Erprobung neuer Medikamente in der Pharmaentwicklung
- Verbesserung der Kosteneffizienz im Gesundheitswesen und Fernüberwachung kritischer Parameter bei Patienten[52].

[52] Vgl. das Anwendungsbeispiel im Unterabschnitt 10.2.3

Echtzeit-Auswertung von komplexen Daten aus wissenschaftlichen Experimenten

Ein hohes Potenzial für den Einsatz von Big Data schlummert in der Wissenschaft sowie in der betrieblichen Forschung und Entwicklung. Ein Beispiel aus dem Forschungszentrum CERN in Genf soll das illustrieren: Während eines Experiments mit dem Teilchenbeschleuniger Large Hadron Collider (LHC) entstehen pro Sekunde 40 Terabytes an Daten.[53]

Produktneuentwicklungen und -verbesserungen

In den Forschungs- und Entwicklungsabteilungen der Unternehmen lassen sich verschiedene Datenbanken mit Kundenbewertungen von Produkten zusammenführen, um Hinweise für das Produktdesign abzuleiten.

Durch diese gezielte Auswertung von Nutzer- und Meinungsforen, sowie Social-Media-Plattformen können systematisch Schwächen und Meinungen zu Produkten und Dienstleistungen ausgewertet werden. Entweder um neue Produktideen zu generieren oder Verbesserungspotentiale an bestehenden Produkten zu identifizieren. Genauso können solche Auswertungen für Sentimentanalysen oder Auswertungen zur Markenwahrnehmung genutzt werden. Hierbei ist natürlich der rechtliche Rahmen und die geltenden Datenschutzrechte zu beachten.

Aus Social-Media-Kanälen wie Facebook oder Twitter sowie aus Blogs und Foren können Unternehmen Ideen für die Weiterentwicklung ihrer Produkte aufnehmen. Umgekehrt wird es kaum sinnvoll sein, Produkte weiterzuentwickeln, die bei den Verbrauchern in solchen Kanälen permanent schlechte Noten erhalten. Wenn Unternehmen kritische Stimmen als Chance und nicht etwa als Niederlage werten, werden sie deutliche Vorteile daraus ziehen, teure Produktionen und überflüssige Marketingkampagnen kurzfristig beenden und stattdessen Verbrauchervorschläge im Sinne des Crowdsourcings für weitere Produkte nutzen.

Social Media für Trendanalysen

Die Auswertung von Social-Media-Kanälen liefert frühe Signale für gesellschaftliche Trends und eröffnet die Chance, Märkte mit genau darauf abgestimmten Produkten zu erschließen[54].

Entwickler sammeln mit Hilfe von virtuellen Kollaborationsseiten oder Ideenmarktplätzen Informationen, die sie mit Partnern teilen und für die Weiterentwicklung von Produkten nutzen. Die kollaborative und parallele Entwicklung sowie die schnelle Umsetzung von Prototypen[55] verkürzt die Time-to-Market und verspricht klare Wettbewerbsvorteile in der Startphase eines Produkts, höhere Absatzchancen und Margen. Erfolgreiche Unternehmen wie Apple oder Google nutzen diese Möglichkeiten virtuos.

Pharmaentwicklung Erprobung neuer Medikamente

In der Pharmaindustrie entstehen bei der Entwicklung und Erprobung neuer Wirkstoffe und Medikamente gigantische Datenmengen, deren Auswertung eine enorme Herausforderung bilden. Big Data eröffnet den Weg zur Aggregation der Daten mehrerer Forschungseinrichtungen und zu ihrer gemeinsamen Auswertung. Rechen- und datenintensive Simulationen führen zu besseren und genaueren Ergebnissen bei der Wirkungsanalyse von Medikamenten und erhöhen die Wahrscheinlichkeit erfolgreicher klinischer Studien. Verringerte F&E-Kosten und verkürzte Time-to-Market stehen zu Buche.[56]

53 Um diese Datenmengen verarbeiten zu können, hat die atomphysikalische Forschungsorganisation CERN sogar eigene Datenbanksysteme entwickelt.

54 Predictive Modeling

55 Rapid Prototyping

56 Im Big-Data-Report von McKinsey (vgl. Fußnote 11) wird eingeschätzt, dass sich die Zeit bis zur Marktreife neuer Medikamente von heute durchschnittlich dreizehn Jahren um bis zu fünf Jahre verkürzen könnte. Das Einsparpotenzial im US-Gesundheitssystem durch die Nutzung von Big-Data-Analysen wird mit jährlich etwa 300 Milliarden Dollar beziffert.



Verbesserung der Kosteneffizienz im Gesundheitswesen

Big Data liefert Ansätze zur Lösung eines bekannten Problems in der Gesundheitsvorsorge. Die Kosten des Gesundheitssystems drohen aus dem Ruder zu laufen. Bisherige Sparansätze sind oftmals isoliert und wenig nachhaltig. Wenn etwa der Zuschuss für teure Medikamente vermindert wird, obwohl deren Wirkung nachweislich die Folgekosten einer Krankheit senken, ist das aus einer langfristigen Perspektive betrachtet wenig sinnvoll. Auch die Vorsorge kommt angesichts des Spardiktats zu kurz. Insgesamt mangelt es an der umfassenden Analyse vorhandener Daten, mit der der Beweis einer langfristig kostendämpfenden Wirkung bestimmter teurer Behandlungen angetreten werden könnte.

Das Gesundheitswesen muss also mit Big-Data-Analysen befähigt werden, eine bessere Gesundheitsvorsorge zu niedrigeren Kosten zu entwickeln. Ein Weg dorthin führt zum Beispiel über komplexe DNA-Analysen, mit denen Ärzte das Auftreten einer Krankheit prognostizieren und proaktiv Gegenmaßnahmen vorschlagen könnten. Auf Basis solcher Analysen ließen sich für Menschen mit sehr ähnlicher DNA-Struktur gruppenspezifisch zugeschnittene Medikamente entwickeln.

7.3 Produktion, Service und Support

Zu den wichtigen Einsatzgebieten von Big Data in der Produktion sowie im Service und Support gehören:

- Produktionsoptimierung mit Maschinen- und Sensordaten
- Produktionsplanung bei Saisonartikeln in Abhängigkeit von vielfältigen Faktoren[57]
- Früherkennung von Produktproblemen im Service durch den Einsatz von Diagnosedaten Vorausschauende Instandhaltung von Maschinen und Anlagen.

Produktionsoptimierung mit Maschinen- und Sensordaten

Mit dem Internet der Dinge oder M2M-Kommunikation können produzierende Unternehmen[58] ihre Fertigungsprozesse optimieren. Dafür erfassen Sensoren an Produkten und entlang von Produktions- und Lieferketten Daten – auch im späteren Betrieb[59]. Die meisten dieser Daten fließen in Echtzeit in Datenbanken ein und werden für Zwecke der Überwachung und Optimierung von Prozessen und wirtschaftlichen Parametern genutzt.

Die Auswertung von Echtzeit-Daten hat in einigen Branchen eine lange Tradition. Das trifft u.a. für die Ölförderung zu; die Erdölverarbeitung setzt Big Data in ihren Raffinerien ein. Mit Daten von Bohrköpfen, seismischen Sensoren oder Telemetrie-Satelliten lassen sich Fehler vermeiden sowie Betriebs- und Wartungskosten senken.

Die verstärkte RFID-Nutzung[60] lässt die Datenflut weiter anschwellen. Bisher bleiben die gespeicherten Daten

57 Hier werden die Methoden eingesetzt, wie sie im Anwendungsbeispiel »(Nº07) Otto – Verbesserung der Absatzprognose« beschrieben werden.

58 Sie speichern heute mehr Daten als jeder andere Industriesektor. 2010 waren es geschätzte zwei Exabyte neuer Daten. Vgl.: IDC Digital Universe Study 2011: Extracting Value form Chaos: http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm (Abruf 28.07.2012)

59 Bei Flugzeugen, Turbinen, Raketen oder großen Maschinen liegen die in wenigen Stunden entstehenden Daten schnell im Terabyte-Bereich.

60 So soll die Zahl von RFID-Tags, die Daten berührungslos erfassen, von zwölf Millionen im Jahr 2011 auf 209 Milliarden im Jahr 2021 steigen. Vgl.: http://www.vdi-nachrichten.com/artikel/Die_Marktsituation_rund_um_RFID/52611/2/GoogleNews (Abruf am 31.08.2012)



meist in isolierten Systemen und dienen ausschließlich einem einzigen Zweck. Immer mehr Produzenten beginnen jedoch, Daten unterschiedlicher Systeme zu verknüpfen und erhöhen damit die Komplexität der Analyse. In der Automobilindustrie und in vielen anderen Industriezweigen nimmt die Fertigungstiefe ab. Die Komponenten von Produkten stammen von vielen Herstellern. In solchen Situationen stellt die übergreifende Qualitätssicherung der Produktion eine besondere Herausforderung dar. Hier können Daten aus CAD-Systemen, dem Engineering, der Fertigung sowie Produktdaten (PLM) übergreifend und zeitnah ausgewertet werden, um die Qualität der Produktion insgesamt sicherzustellen.
Die gemeinsame Nutzung der Daten kann das PLM verbessern. Kundenmeinungen fließen in das Design der Produkte ein und Open-Innovation-Ansätze werden praktiziert.

In Kombination mit Digital-Factory-Simulationen lässt Big-Data-Analytics Schwachstellen von Produktionsprozessen erkennen und leistet Beiträge zu deren substanzieller Verbesserung.

Früherkennung von Produktproblemen im Service durch den Einsatz von Diagnosedaten

Komplexe technische Geräte produzieren große Mengen an Daten im Betrieb, die nur unzureichend ausgewertet werden. Erst wenn das »Kind in den Brunnen gefallen« ist, analysieren Hersteller und Betreiber, auf welche Fehler die Störung zurückzuführen ist. Diese nachträgliche Fehlerbetrachtung hilft zwar, auf Basis dieser Informationen nachfolgende Produktreihen in der Fertigung zu verbessern, den aktuell von einem Fehler betroffenen Kunden hilft es aber wenig. Mit Big Data lassen sich Produkte auch im laufenden Betrieb zuverlässig überwachen und intelligente Diagnosen mit Trendanalysen erstellen, die präventiv wirken können. Dazu werden aktuelle Produkt- und Sensorinformationen mit Informationen aus dem Service oder früheren Fehlern korreliert und ausgewertet. So lassen sich Kundenservicemodelle verbessern und der Vertrieb optimieren.

Im Falle eines Defektes lassen sich die Fehlerquelle unverzüglich feststellen - eventuell remote beheben - und Korrekturen im Fertigungsprozess einleiten. Big Data ermöglicht damit präventive Wartung (Predictive Maintenance), indem über Sensoren sämtliche Informationen über den Zustand von Anlagen und relevante Umgebungsdaten wie Raumtemperaturen oder Luftfeuchtigkeit erfasst und ausgewertet werden. Maschinenbauer erkennen Störungen auf diese Weise frühzeitig und verhindern ungeplante Stillstände, indem sie mit Big-Data-Analyseverfahren Zusammenhänge zwischen Indikatoren aufdecken, die nicht offensichtlich sind. So lassen sich Störungen aufdecken, bevor sie Schaden anrichten, was Stillstandzeiten verringert und Wartungskosten spart.

7.4 Distribution und Logistik

Die wichtigsten Einsatzgebiete in Distribution und Logistik:

- Optimierung von Lieferketten (Supply Chain Management)
- Nutzung von Verkehrstelematik zur Verminderung von Stillstandszeiten bei LKW-Transporten
- Optimierung bei der Mauterhebung

Optimierung von Lieferketten (Supply Chain Management)

Unternehmen müssen zur Optimierung ihrer Produktionsmengen und Lieferketten zeitnah hohe Volumina vielfältiger Daten verarbeiten. Neben Daten der eigenen Kapazitäten an den Produktionsstandorten sind das Daten der Auftragsfertiger, der Zwischenlager und der Logistikpartner sowie Prognosen künftiger Absatzmengen. Dieses Problem wird komplexer, wenn ein Hersteller eine Vielzahl von Produkten berücksichtigen muss. Eine zeitnahe und verteilte Auswertung der Daten, die auch bei den Partnern liegen können, ist in dieser Situation erfolgsentscheidend.



Nutzung von Verkehrstelematik zur Verminderung der Stillstandszeiten bei LKW-Transporten

Transportlogistiker agieren in einem eng umkämpften Markt und sind mit mehreren Herausforderungen konfrontiert. Die Verkehrsdichte in Deutschland nimmt weiter zu, kilometerlange Staus verzögern Transporte. Steigende Kosten für Fahrzeuge, Betrieb, Wartung und Benzin können nur teilweise an Kunden weiterberechnet werden. Spediteure suchen daher nach Hebeln, die Kostenspirale zu stoppen, unter anderem auf dem Wege einer stärkeren Vernetzung der Fahrzeuge mit der Außenwelt. Immer mehr LKW werden mit Sensoren und Steuerungsmodulen ausgestattet, die Fahrzeugdaten wie den Benzinverbrauch, den Zustand von Verschleißteilen oder Positionsdaten erfassen und in Datenbanken übertragen. Mit diesen Daten können Disponenten zeitnah Transporte planen, gegebenenfalls Routen und Beladung ändern und die nach wie vor hohe Zahl von Leerfahrten verringern. Mit den Fahrzeugdaten lassen sich Werkstattaufenthalte dem tatsächlichen Bedarf anpassen und Stillstandzeiten minimieren. Umgekehrt fließen Daten in die Cockpit-Systeme zurück und passen auf Basis der aktuellen Verkehrslage automatisch Routen an, um Transportzeiten zu minimieren.

Optimierung bei der Mauterhebung

Ein weiteres Szenario basiert ebenfalls auf Verkehrs- und Logistikdaten. Die Wirtschaftsleistung einer Region oder eines Landes lässt sich derzeit nur rückblickend erfassen, wenn Daten der Unternehmen und Verbände vorliegen. Mit Big Data wäre es möglich, aktuelle Daten wie das regionale Verkehrsaufkommen aus Navigationssystemen und Daten der Mauterhebung zu kombinieren mit Produktions- und Transportinformationen der Unternehmen. So ließe sich das aktuelle Wirtschaftsgeschehen sowie Verkehrstrends in Echtzeit erkennen.

Fahrerlose Fahrzeuge

Inzwischen gibt es auch Prototypen von Autos, die mithilfe von Sensoren durch den Stadtverkehr fahren, ohne dass ein menschlicher Fahrer lenken muss. Das Leitsystem findet durch Datenanalyse die beste, schnellste, staufreie Route zum Ziel. Fängt es an zu regnen, fährt der Wagen automatisch langsamer, weil sich der Bremsweg verlängert. Fahrer und Leitsysteme haben Zugriff auf ein riesiges Netzwerk von Sensordaten und optimieren damit den Verkehrsfluss. Erste Testprojekte sollen zudem aufzeigen, wie sich Sensordaten aller Fahrzeuge in Echtzeit auswerten lassen und sich wichtige Informationen wieder zurückspielen lassen. Dies könnte Verkehrsunfälle verhindern. Wenn zum Beispiel vernetzte Fahrzeuge die Information über Stauenden via Sensoren an nachfolgende Fahrzeuge senden könnten, ließen sich gefährliche Auffahrunfälle deutlich verringern.

Wenn solche Szenarien umgesetzt werden, entstehen enorme Datenmengen, die aus den Fahrzeugen in Datenbanken zusammenlaufen, dort mit anderen Daten korreliert werden müssen und dann teilweise in Echtzeit wieder verteilt werden. Solche Massendaten lassen sich nur mit Big-Data-Analysemethoden bewältigen.



7.5 Finanz- und Risiko-Controlling

Zu den wichtigsten Einsatzgebiete im Bereich Finanz- und Risiko-Controlling gehören:

- Echtzeit-Reaktionen auf Geschäftsinformationen
- Simulationen, Vorhersagen und Szenarienbildung
- Manipulationsprävention und Betrugserkennung
- Risikocontrolling in Echtzeit
- Erkennen von Kreditrisikofaktoren.

Echtzeit-Reaktionen auf Geschäftsinformationen

Durch neue Big-Data-Technologien können die unterschiedlichsten Unternehmensinformationen schnell zusammengeführt und für Entscheidungen genutzt werden. Es stehen weit mehr Daten, Fakten und Beobachtungen als je zuvor zur Verfügung. Gleichzeitig ermöglichen es neue Herangehensweisen in Soft- und Hardware, diese Informationen analytisch so aufzubereiten, dass sie im Geschäftsprozess zur Verfügung stehen. Damit können beispielsweise kundenindividuelle Rabatte direkt im Callcenter oder auf der Website gewährt werden, die noch ganz aktuelle Informationen in die Berechnung der Rentabilität einbeziehen. Im Gegensatz zu klassischen Business-Intelligence-Ansätzen, in denen vornehmlich formatierte Berichte auf starren Datenmodellen zum Einsatz kommen, erlauben Big-Data-Ansätze in weit stärkerem Maß als zuvor die gezielte Beantwortung von Ad-hoc-Fragestellungen. So können Informationen aus verschiedenen Abteilungen und Systemen für eine gemeinsame Analyse verwendet werden, ohne dass zuvor ein klassisches Datenmodell aufgebaut werden muss. Die Fachexperten gewinnen damit ein Mehr an Freiheit, das ihnen ein iteratives Vorgehen und selbständiges Arbeiten ohne Limitationen durch eingeschränkt verfügbare IT-Ressourcen ermöglicht.

Simulationen, Vorhersagen und Szenarienbildung

Die Fähigkeit zur Entwicklung von Prognosen rückt – in Ergänzung zu klassischen Berichtsfunktionen – ebenfalls in den Fokus. Zur Erstellung von aussagekräftigen Vorhersagemodellen werden große Datenmengen und hohe Rechenleistung benötigt. Dabei steigt die Aussagekraft, etwa über die zu erwartenden Absatzzahlen im nächsten Monat, mit der Menge der zur Verfügung stehenden Informationen aus der Vergangenheit. Wenn bisher die Berechnung von Vorhersagemodellen aufgrund der Komplexität der Fragestellung oder der Menge der anfallenden Daten ein aufwändiger Prozess mit langen Laufzeiten war, sind künftig »Was wäre, wenn«-Analysen möglich. Statt nur einziges Modell zu erstellen, können sehr viele verschiedene berechnet und in ihren Auswirkungen verglichen werden. Die Simulation verschiedener möglicher Entwicklungen unterstützt die Entscheidung für die beste Alternative. Das hilft in der Planung der einzusetzenden Ressourcen in der Produktion genauso wie in der Optimierung der jeweiligen Logistikketten im Absatz.

Manipulationsprävention und Betrugserkennung

Im Bereich Betrugserkennung können nun wesentlich mehr Verhaltensmuster beobachtet, erkannt und in Echtzeit bearbeitet werden. Auch bei Millionen von Transaktionen, deren Auswirkungen erst in der Kombination verschiedener Daten aus verschiedenen Quellen sichtbar werden, gibt es nun Lösungen, die etwa die missbräuchliche Verwendung von Kreditkarten erkennen. Technisch gesehen geht es dabei auf der einen Seite wiederum um die Entwicklung von Modellen, auf der anderen Seite um eine Echtzeitanbindung der operativen Systeme an analytische Anwendungen. Dieses »Scoring« prüft jede einzelne Transaktion über ein möglichst präzises und dennoch einfach zu handhabendes Modell auf Auffälligkeiten und erlaubt das sofortige Ableiten von Aktionen – etwa einer Sperrung oder einer nachträglichen Detailprüfung. Im Ergebnis werden wesentlich mehr Betrugsfälle erkannt, es können Gegenmaßnahmen eingeleitet und letztlich der Schaden minimiert werden. Damit sinken die Kosten für den Verbraucher und das Vertrauen in sichere Zahlungssysteme wird aufrechterhalten.

Risikocontrolling in Echtzeit

Erst aus einer großen Fülle von Einzelinformationen lassen sich Aussagen über ein unternehmenskritisches Risiko ableiten. Im Bankbereich etwa müssen sowohl Marktzahlen wie auch unternehmensinterne Entwicklungen in Gänze überwacht und in die Risikoallokation einbezogen werden. Die dabei entstehenden Datenmengen übersteigen die Kapazität heutiger Systeme vor allem in Bezug auf die Verarbeitungsgeschwindigkeit. Mit Big-Data-Technologien wird es möglich, komplexe Value-at-Risk-Berechnungen in sehr viel kürzerer Zeit durchzuführen und damit auf Ereignisse wie fallende Kurse sehr viel schneller zu reagieren. Das zur Verfügung stehende Eigenkapital einer Bank kann damit effizienter eingesetzt und riskante Auswirkungen auf die eigenen Risikopositionen können so rechtzeitig erkannt werden, dass die Möglichkeit für die Einleitung wirksamer Gegenmaßnahmen besteht.

Erkennen von Kreditrisikofaktoren

Einer der Auslöser der Finanzkrise von 2008 war das Platzen der Immobilienblase in den USA. Kredite, die zur Anschaffung von Immobilien gewährt wurde, konnten nicht mehr bedient werden. Es war und ist eine der wichtigsten Aufgaben von Banken, Kredite aufgrund zugrundeliegender Risiken zu bewerten und letztlich zu bepreisen. Mit Big-Data-Ansätzen lassen sich nun wesentlich mehr Einflussfaktoren aus wesentlich mehr Quellen einbeziehen. Die Fähigkeit, ein passendes und gültiges Modell über die Ausfallwahrscheinlichkeit zu ermitteln, wächst. Damit können deutlich besser als vorher kundenindividuelle Kreditangebote unterbreitet und systemrelevante Klumpenrisiken – »zu viele faule Immobilienkredite« – erkannt werden.



8 Big Data und Datenschutz

Während in den USA der Blick mehr auf die großen Chancen von Big Data gerichtet ist, werden in Deutschland eher die Risiken von Big Data betont und Befürchtungen vor unkontrollierter Überwachung thematisiert.

Nach deutschem Datenschutzrecht sind Big-Data-Methoden in einer ganzen Reihe von Fällen zulässig.

Die rechtlichen Herausforderungen bestehen darin, in Vertragsverhältnissen zu beurteilen, welche Datenverarbeitung erforderlich ist, für wirksame Einwilligungen zu sorgen und taugliche Verfahren zum Privacy-Preserving Data Mining anzuwenden. Vor allem ist es wichtig, die rechtliche Zulässigkeit bereits bei der Entwicklung einer Big-Data-Anwendung zu prüfen. Die rechtliche Zulässigkeit hängt nämlich stark vom Design des Verfahrens ab. In der Anfangsphase der Entwicklung lässt sich das einfacher ändern als später, wenn ein Verfahren bereits eingeführt ist.

Big-Data-Verfahren nach deutschem Recht unter bestimmten Voraussetzungen zulässig

Das Potential von Big Data ist enorm: Der technische Fortschritt macht es möglich, immer größere Datenmengen in immer kürzerer Zeit auszuwerten. In Deutschland wird Big Data aber nur zu einem Erfolg werden, wenn sich die Verfahren mit dem deutschen Datenschutzrecht in Einklang bringen lassen.

Das US-amerikanische Recht gibt Unternehmen sehr weitreichende Rechte, Daten zu verarbeiten. Dagegen ist das deutsche Recht restriktiv. Ein wichtiger Grundsatz des Bundesdatenschutzgesetzes (BDSG) ist das Verbotsprinzip: Personenbezogene Daten dürfen nur erhoben, verarbeitet oder genutzt werden, wenn der Betroffene eingewilligt hat oder wenn eine Rechtsvorschrift dies ausdrücklich erlaubt. Daten dürfen nur für den Zweck genutzt werden, für den sie erhoben worden sind (Zweckbindung). Es sollen möglichst wenige personenbezogene Daten verarbeitet werden (Datensparsamkeit).

Dies bedeutet aber keineswegs, dass Big-Data-Verfahren nach deutschem Recht unzulässig wären. Einige Beispiele zeigen dies.

Wichtige Bereiche sind

- die Datenverarbeitung in Vertragsverhältnissen,
- die Verarbeitung auf Grund einer Einwilligung des Betroffenen und
- das Privacy-Preserving Data Mining.

Fraud Detection und Kredit-Scoring

Eines der größten Probleme von Online-Zahlungsdiensten und Kreditkartenunternehmen ist der Missbrauch durch Betrüger. PayPal hat zur Abwehr die Softwareanwendung Igor entwickelt und konnte sich damit zu einem der erfolgreichsten Zahlungsdienstleister entwickeln. Visa und MasterCard setzen ebenfalls leistungsfähige Software ein, um verdächtige Zahlungen zu ermitteln.

Auch nach deutschem Datenschutzrecht ist der Einsatz von Fraud-Detection-Anwendungen grundsätzlich zulässig, denn die Verarbeitung der personenbezogenen Daten ist erforderlich, damit der Zahlungsdienstleister den Vertrag durchführen kann. Der Dienstleister muss seine Kunden davor schützen, dass Diebe oder Betrüger Kredit- oder EC-Karten missbrauchen.

Big Data verbessert die Möglichkeiten, Muster im Zahlungsverhalten zu finden und bei Auffälligkeiten sofort zu reagieren. Je leistungsfähiger die Analysemethoden

sind, desto besser für den Kunden: Betrug wird dadurch erschwert, falsch positive Fälle, also die Nichtausführung von Zahlungen, die in Wirklichkeit von dem Berechtigten veranlasst wurden, werden reduziert. Zahlungsdienstleister, die Fraud Detection einsetzen, müssen natürlich dafür sorgen, dass die Daten nur für diesen Zweck verwendet werden und nicht in falsche Hände geraten können.

Ein weiteres Beispiel ist das Kredit-Scoring. Eine Bank muss vor der Entscheidung über einen Kredit die Kreditwürdigkeit ihres Kunden bewerten. Hierzu setzen Banken u.a. auch Scoring-Verfahren ein. Bei den üblichen Scoring-Verfahren werden Angaben zu Beruf, Einkommen, Vermögen, bisheriges Zahlungsverhalten, etc. ausgewertet, insgesamt relativ wenige Parameter. Demgegenüber wenden manche amerikanische Kreditanbieter mit Erfolg Methoden an, die tausende Indikatoren nutzen.

Auch nach deutschem Recht wäre dies grundsätzlich möglich. Die Bank darf lediglich Angaben zur Staatsangehörigkeit und besonders schutzwürdige Daten des Kunden, wie z. B. Angaben zur Gesundheit, nicht für das Kredit-Scoring verwenden. Andere Daten darf die Bank nutzen, vorausgesetzt sie sind nach wissenschaftlich anerkannten mathematisch-statistischen Verfahren nachweisbar für die Risikoermittlung geeignet. Big Data kann helfen, diesen Nachweis zu führen. Verbesserte Analysemethoden sind vorteilhaft für Banken, aber genauso auch für Kunden.

Jedes Unternehmen verarbeitet personenbezogene Daten im Vorfeld eines Vertragsschlusses oder um geschlossene Verträge abzuwickeln. Das ist datenschutzrechtlich auch zulässig. Hier kann auch Big Data zum Einsatz kommen. Die rechtliche Herausforderung besteht darin, in jedem Einzelfall zu prüfen und zu begründen, warum die Datennutzung für den Abschluss oder die Abwicklung des Vertrages erforderlich ist.

Kundenbindungssysteme

Tesco, die große britische Supermarktkette, wertet Daten aus Einkäufen aus, um Kunden gezielt Coupons mit maßgeschneiderten Angeboten zu übersenden. Auch nach deutschem Datenschutzrecht ist es zulässig, Kundendatenbanken zu verwenden, um für Produkte zu werben. Werbung ist zwar nicht mehr für die Erfüllung eines bestehenden Vertrages erforderlich, das Unternehmen hat aber ein berechtigtes Interesse daran, Kunden über sein Angebot zu informieren. Damit ist der Einsatz von CRM-Anwendungen zulässig.

Will ein Unternehmen allerdings mithilfe von Data Mining Kundenprofile anlegen, die eine zielgerichtete Werbung ermöglichen, so erfordert dies die Einwilligung des Kunden in die Auswertung seiner Daten. Big Data erleichtert dies. Der Kunde wird seine Einwilligung nämlich nur erteilen, wenn er sich hiervon einen Nutzen verspricht. Gerade wenn es dem Unternehmen durch Big-Data-Verfahren gelingt, den Bedarf seines Kunden besser zu erkennen und die Angebote hierauf zuzuschneiden, wird ein Kunde hierin einen Vorteil sehen.

Die rechtliche Herausforderung besteht darin, eine rechtswirksame Einwilligung des Kunden zu erhalten. Die Anforderungen sind hier hoch. Die Einwilligung, die Daten zu verarbeiten, kann in Allgemeinen Geschäftsbedingungen erteilt werden, sofern sie besonders hervorgehoben ist. Der Kunde muss dann den Text streichen, wenn er nicht einverstanden ist (Opt-out). Will das Unternehmen aber Coupons oder andere Produktwerbung per E–Mail übersenden, muss der Kunde ausdrücklich zustimmen, z. B. indem er dies gesondert ankreuzt (Opt-in). Schwierig ist es, die Einwilligung so weit zu fassen, dass sie auch alle Auswertungen der Daten abdeckt, sie aber gleichzeitig ausreichend zu präzisieren – sonst ist sie unwirksam. Bloße allgemeine Wendungen wie »Zum Zwecke der Werbung« reichen nicht aus, wenn Data Mining betrieben werden soll.

Privacy-Preserving Data Mining

Das Datenschutzrecht gilt nur für personenbezogene Daten. Dazu gehören beispielsweise Name, Geburtsdatum, Anschrift, E-Mail-Adresse. Daten ohne Personenbezug werden hiervon nicht erfasst, also technische Daten wie zum Beispiel ausgewertete Maschinendaten, Gerätedaten für Service- und Support- Zwecke oder andere technische Daten im Sektor Forschung und Produktentwicklung.

Ein großes deutsches Marktforschungsunternehmen erfasst das Einkaufsverhalten von 15.000 Haushalten im Detail, verarbeitet dann aber ausschließlich anonymisierte Daten. Dieses Privacy-Preserving Data Mining erhöht nicht nur die Bereitschaft der Verbraucher, Daten preiszugeben, sondern es ist auch datenschutzrechtlich erheblich einfacher, die Daten zu verarbeiten.

Anonymisieren bedeutet, die Daten so zu verändern, dass sich nicht mehr bestimmen lässt, welcher Person sie zuzuordnen sind. Methoden sind das Löschen aller Identifikationsmerkmale und das Aggregieren von Merkmalen, hier wird beispielsweise die Adresse durch eine Gebietsangabe ersetzt oder das Geburtsdatum durch das Geburtsjahr.

Dabei ist es nicht einmal erforderlich, die personenbezogenen Ausgangsdaten zu löschen. Erstellt man aus den Rohdaten einen abgeleiteten anonymisierten Datenbestand und leitet ihn einem Dritten zu, so sind die Daten für diesen nicht personenbezogen, vorausgesetzt der Dritte hat keine Möglichkeit, den Personenbezug herzustellen.

Hierzu muss man vertragliche Vereinbarungen treffen, die ein Zusammenführen ausschließen. Außerdem muss man prüfen, dass der Dritte nicht über Zusatzwissen verfügt, um die Daten zu deanonymisieren und dass er dieses Zusatzwissen auch nicht erlangen kann.

Eine weitere Methode ist es, mit Pseudonymen zu arbeiten. Bei der Pseudonymisierung werden der Name des Betroffenen und andere Identifikationsmerkmale durch Kennzeichen ersetzt. Das hat den Vorteil, dass Daten weiter als Profil verarbeitet werden können, was beispielsweise in der medizinischen Forschung wichtig ist, wenn Krankengeschichten ausgewertet und Längsschnittstudien erstellt werden sollen.

Es gibt verschiedene Grade der Pseudonymisierung. Eine Methode ist die Einweg-Pseudonymisierung z. B. durch die Zuordnung von Hash-Werten, die keine Umkehrung zulässt. Dann verlieren die Daten generell ihren Charakter als personenbezogene Daten, werden also anonyme Daten. Es gibt auch andere Möglichkeiten: Erstellt der Inhaber einer Datenbank Pseudonyme und fertigt er hierbei eine Referenztabelle an, die er in seinem Besitz hält, so sind die Daten für ihn weiterhin personenbezogen. Wenn er die pseudonymisierten Daten an einen Dritten weitergibt, sind sie dort aber nicht mehr personenbezogen, sofern die Weitergabe der Referenztabelle durch vertragliche Vereinbarungen ausgeschlossen und auch keine Deanonymisierung möglich ist. Der Empfänger der Daten kann die Daten dann als nicht personenbezogene Daten verarbeiten.

Privacy-Preserving Data Mining ist nicht nur für den Bereich der Marktforschung geeignet, sondern auch für das Webtracking oder auf Gebieten wie der medizinischen Forschung, wo besonders schutzbedürftige Daten verarbeitet werden. Allerdings ist der Einsatz von Privacy-Preserving Data Mining rechtlich anspruchsvoll. Bei einigen Daten, z. B. dynamischen IP-Adressen, ist es juristisch umstritten, ob sie personenbezogen sind. Bei anonymisierten Daten muss das Risiko bewertet werden, dass sie mit anderen Daten zusammengeführt werden. Wenn der Aufwand an Zeit, Kosten und Arbeitskraft unverhältnismäßig hoch wäre, ist dieses Risiko rechtlich unbeachtlich.

Entwicklung einer Big-Data-Anwendungen - rechtliche Zulässigkeit im Vorfeld klären

In den USA ist der Blick mehr auf die großen Chancen von Big Data gerichtet, wie auf die Aufdeckung von Nebenwirkungen von Medikamenten, wie z. B. des Rheumamittels Vioxx im Jahre 2004. Eine Menschenrechtsorganisation setzt Big-Data-Methoden ein, um zu untersuchen, ob die amerikanische Justiz Minderheiten benachteiligt. In Deutschland werden demgegenüber eher die Risiken von Big Data betont und eine unkontrollierte Überwachung befürchtet. Auch nach deutschem Datenschutzrecht sind Big-Data-Methoden aber in einer ganzen Reihe von Fällen zulässig.

Die rechtlichen Herausforderungen bestehen darin, in Vertragsverhältnissen zu beurteilen, welche Datenverarbeitung erforderlich ist, für wirksame Einwilligungen zu sorgen und taugliche Verfahren zum Privacy-Preserving Data Mining anzuwenden. Vor allem ist es wichtig, die rechtliche Zulässigkeit bereits bei der Entwicklung einer Big-Data-Anwendung zu prüfen. Die rechtliche Zulässigkeit hängt nämlich stark vom Design des Verfahrens ab. In der Anfangsphase der Entwicklung lässt sich das einfacher ändern als später, wenn ein Verfahren bereits eingeführt ist.

9 Big Data – Marktentwicklung in wichtigen Regionen[61]

Big Data steht trotz der frühen Marktphase schon für ein hoch relevantes IT-Marktsegment, in dem im Jahr 2011 global Umsätze in der Größenordnung von 3,3 Milliarden Euro getätigt wurden. Junge, allein auf Big Data ausgerichtete Technologiefirmen erzielten mit 270 Millionen Euro rund 8% der globalen Umsätze.

- In der derzeitigen Marktphase stehen Auf- und Ausbauinvestitionen auf der Infrastrukturseite bei den Anwendern im Vordergrund, um die Grundlage für Big-Data-basierte Prozesse und Geschäftsmodelle zu legen.
- Der globale Markt wird von derzeit (2012) rund 4,5 Milliarden Euro auf zukünftig (2016) 15,7 Milliarden Euro anwachsen (CAGR 36%).
- Obwohl deutsche Unternehmen deutlich bedächtiger in das Thema einsteigen, wird Deutschland eine europaweite Führungsrolle zum Thema Big Data einnehmen. Die Notwendigkeit, als hoch wettbewerbsfähige und exportorientierte Volkswirtschaft seine Produktions-, Logistik- und Vertriebsketten weltweit optimiert zu planen und zu steuern, wird deutsche Unternehmen zu »Big Data Champions« machen.
- In der derzeit frühen Marktphase sind Unternehmen aus der Internet- und eCommerce- und Werbebranche die Vorreiter beim Einsatz von Big Data.



Abbildung 12: Globaler Markt für Big Data in drei Kennziffern; Quelle: Experton Group 2012

Die Zeit ist klar vorbei, Big Data als Marketing- und Trendbegriff abzuqualifizieren. Das Marktsegment entwickelt sich auf globaler Ebene sehr dynamisch (vgl. Abbildung 12) und zugleich professionell: Die mit Big Data verbundenen Problemstellungen und Business-Herausforderungen sind mittlerweile klar umrissen.

So stellt das hochdynamische Datenwachstum in den Unternehmen deren IT-Entscheider vor die Herausforderung, die IT- und Netzwerk-Infrastruktur aufzurüsten und anzupassen.

Investitionsbereiche

Big-Data-Lösungen haben somit eine starke Investitionskomponente in den Bereichen Hardware und Infrastruktur. Über neue Anwendungen[62] und Delivery-Konzepte wie SaaS und IaaS steigt vor allem das Datenaufkommen in den Unternehmensnetzwerken und den Schnittstellen zur Public Cloud, was den Bedarf an Bandbreite und zugehörigen Acceleration-Services antreibt.

---

61 Die Zahlen und Einschätzungen in diesem Kapitel fußen auf einer Studie der Autoren Dr. Carlo Velten und Steve Janata: Datenexplosion in der Unternehmens-IT: Wie Big Data das Business und die IT verändert (Eine Studie der Experton Group AG im Auftrag der BT (Germany) GmbH & Co. oHG), Dr. Carlo Velten, Steve Janata, Mai 2012

62 z. B. Videoconferencing, Filesharing, Collaboration



Ein weiterer zentraler Investitionsbereich sind die neuen Datenbank- und Analytics-Technologien, die derzeit auch die Aufmerksamkeit der Investoren auf sich ziehen. So wurden in den letzten 24 Monaten über 30 neue Big-Data-Startups mit Kapital ausgestattet, um ihre Technologien und Services zur Marktreife zu entwickeln und global zu vertreiben.

Diese Firmen bilden den Kern der spezialisierten Big-Data-Anbieter, deren Unternehmensstrategie und Portfolio ausschließlich auf dieses Thema ausgerichtet sind. Vielfach agieren diese als Technologie- und Know-how-Partner der etablierten und globalen Technologie- und IT-Serviceanbieter, die im Rahmen der oft komplexen und investitionsintensiven Projekte die Verantwortung übernehmen. Immerhin sind die spezialisierten Big-Data-Anbieter für rund 8% der globalen Ausgaben und Investitionen verantwortlich[63] und belegen damit die These, dass Big Data ein neues Marktsegment bildet und nicht ein recyceltes »BI« darstellt, wie mitunter vermutet wird.

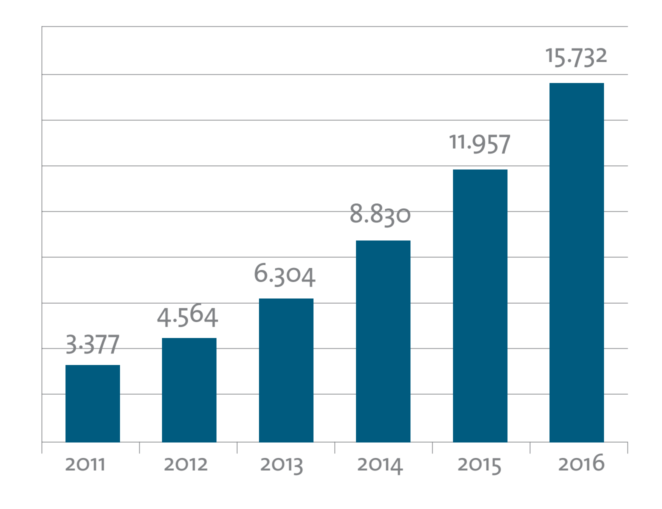


Abbildung 13: Entwicklung des globalen Big-Data-Marktes 2011-2016 in Mio. €; Quelle: Experton Group 2012

Globale Marktentwicklung

Die aktuellen Marktzahlen (vgl. Abbildung 13) zeigen, dass im Jahr 2011 weltweit schon über 3,3 Milliarden Euro in Big-Data-Lösungen und -Services investiert wurden. Bis zum Jahresende 2012 werden sich die Ausgaben und Investitionen auf 4,5 Milliarden Euro erhöhen. Das mittlere Jahreswachstum (CAGR) beträgt über die kommenden 5 Jahre rund 36%. Der Markt für Big Data zählt somit zu den wachstumsstärksten Segmenten des gesamten IT-Marktes. Den Wachstumspeak wird in den Jahren 2013 (38%) und 2014 (40%) erwartet, getrieben durch die initialen Infrastrukturinvestitionen. In den Folgejahren wird sich das Wachstum global wieder leicht abflachen 2016 (31%). Dies ergibt sich einerseits durch den Preisverfall auf der Technologieseite, als andererseits durch den verstärkten Einsatz von internem Know-how in den Unternehmen. Diese sind dann in der Lage, Big-Data-Projekte auch mit eigenen Mitarbeitern umzusetzen, während diese Kompetenzen in den nächsten 2-4 Jahren sehr teuer zugekauft werden müssen.

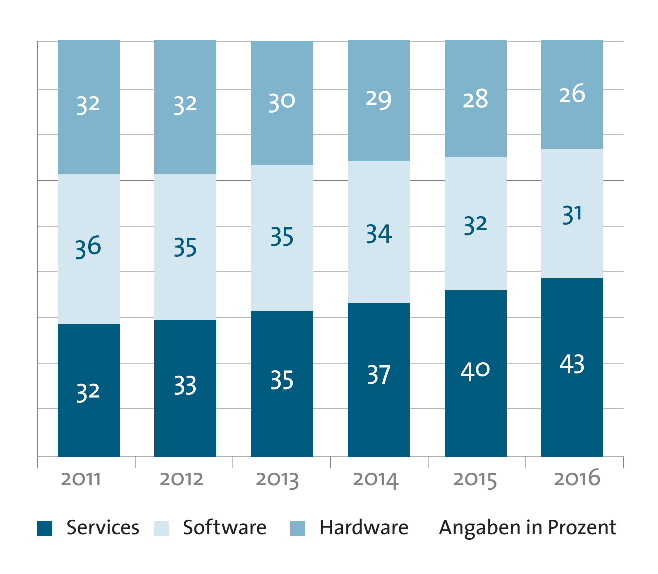


Abbildung 14: Struktur des globalen Big-Data-Marktes 2011-2016 in Mio. €; Quelle: Experton Group 2012

[63] insgesamt € 270 Mio. im Jahr 2011



Marktentwicklung nach Umsatzkategorien

Betrachtet man die Aufteilung der Big-Data-Investitionen nach Hardware, Software und Services (vgl. Abbildung 14), fällt auf, dass die derzeitige Marktphase eine leichte Übergewichtung auf Infrastruktur- und Softwareseite mit sich bringt. Die Auf- und Ausrüstung auf Server-/ Storage- und Netzwerkseite sowie die Lizensierung neuer Software-Lösungen und Tools ist in den meisten Unternehmen allerdings die Grundvoraussetzung, um Big-Data-Projekte zu realisieren. Zukünftig wird mit wachsender Marktreife wieder eine Verschiebung in Richtung Services stattfinden (43% aller Ausgaben in 2016). Elementare Voraussetzung ist hier der Aufbau von Vertrauen in die Big-Data-Dienstleister in Bezug auf Integrität und Leistungsfähigkeit.

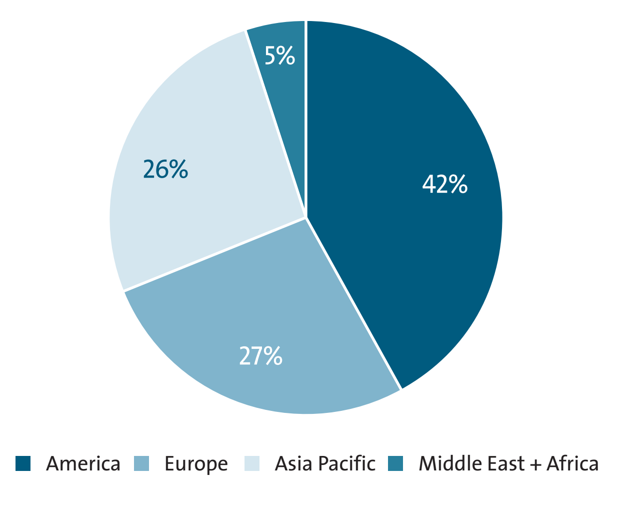


Abbildung 15: Struktur des globalen Big-Data-Marktes nach Regionen 2012; Quelle: Experton Group 2012

Regionalstruktur des Big-Data-Umsatzes

Erwartungsgemäß entfällt in dieser frühen Marktphase noch ein wesentlicher Teil der Umsätze auf den amerikanischen Raum (42% in 2012, vgl. Abbildung 15). So ist in den USA der Einsatz von Big Data mittlerweile weit verbreitet. So sind die USA derzeit noch der zentrale Big-Data-Entwicklungsstandort. Dies lässt sich mit der Anzahl an Firmengründungen und -Finanzierungen, den F&E-Ausgaben sowie anhand der Tatsache erklären, dass diejenigen Branchen, die federführend beim Big-Data-Einsatz sind, in den USA einen großen volkswirtschaftlichen Stellenwert einnehmen (z. B. eCommerce, Online-Werbung, Social Media & Gaming etc.).

Aber auch Europa nimmt aus globaler Perspektive einen zentralen Platz ein. So werden rund 27% der globalen Umsätze mit Big-Data-Lösungen in Europa generiert. Bedenkt man, dass in den USA Regierungsbehörden und Nachrichtendienste für einen wesentlichen Teil der Umsätze stehen, relativiert sich das Bild von der eindeutigen Vorreiterrolle der USA.

Derzeit entfallen rund ein Viertel der Investitionen und Ausgaben auf die Region Asia-Pacific. Dies liegt einerseits am hohen Einsatzgrad von Big Data in Japan[64] sowie den weitgehenden Befugnissen chinesischer (und anderer) Behörden und Unternehmen im Umgang mit personen- und firmenbezogenen Daten. Das starke Wachstum der Internet-, eCommerce- und Gaming-Industrie ist ein weiterer Treiber der Big-Data-Initiativen im asiatischen Raum.

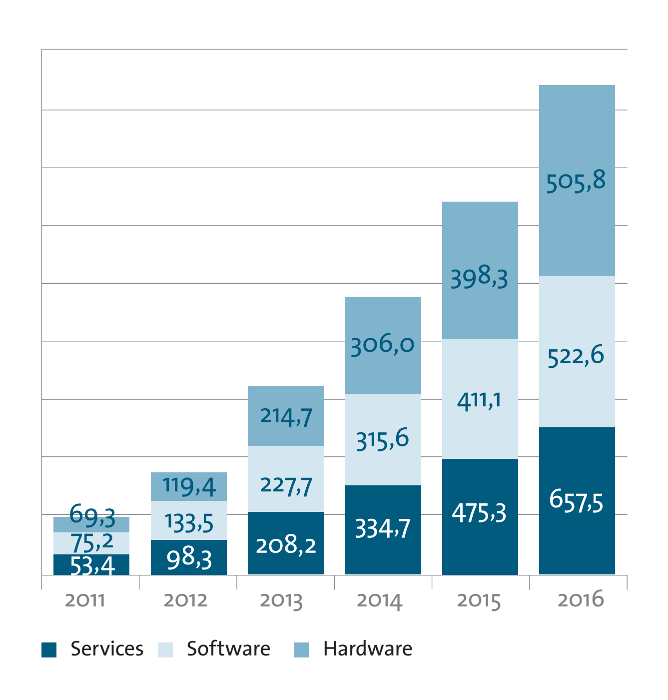


Abbildung 16: Deutscher Big-Data-Markt 2011-2016 nach Marktsegmenten in Mio. €; Quelle: Experton Group 2012

64 intensive Prozessoptimierung, Big Data als Grundlage für »Kaizen«



Deutscher Markt für Big Data – eine Aufholjagd?

Deutschland nimmt zum Thema Big Data eine Sonderrolle ein (vgl. Abbildung 16). Während andere Länder beim Big-Data-Einsatz voranpreschen, verhalten sich deutsche Unternehmen eher wie Mittel- und Langstreckenläufer. Sie agieren derzeit noch recht verhalten und befinden sich mehrheitlich noch in der Sondierungs- und Analysephase. So entfallen auf deutsche Unternehmen derzeit nur ein Fünftel der Big-Data-Umsätze in Europa. Dies wird sich in den kommenden 5 Jahren dramatisch verändern. Es kann davon ausgegangen werden, dass Deutschland im Jahr 2016 für rund die Hälfte der europäischen Big-Data-Umsätze stehen wird. Die Gründe für diese Aufholjagd sind vielfältig. Entscheidend wird die ausgeprägt wettbewerbsfähige Situation der deutschen Unternehmen im europäischen Vergleich sowie ihre Exportabhängigkeit sein. So müssen deutsche Unternehmen die Planung und Steuerung ihrer globalen Produktions- und Lieferketten immer weiter optimieren. Und hier liefert Big Data unterschiedliche Lösungsansätze. Auch wird die Ingenieurs- und Prozessorientierung in der Unternehmensführung und -organisation für einen breiten und vertieften Einsatz von Big Data in deutschen Unternehmen sorgen. Es wird erwartet, dass der Markt für Big-Data-Lösungen in Deutschland von derzeit € 198 Millionen auf rund € 1,6 Milliarden in 2016 anwächst. Dies würde von der derzeit geringen Basis ein jährliches Wachstum von rund 48% bedeuten.

Big Data in volkswirtschaftlichen Bereichen

Die Tabelle 9 stellt den Versuch einer zusammenfassenden Abschätzung dar, in welchen volkswirtschaftlichen Bereichen besonders intensive Veränderungen zu erwarten sind.

Branche	Datenintensität Heute – 2012 (1=niedrig/10=hoch)	Datenintensität Zukünftig – 2020 (1=niedrig/10=hoch)	Datenwachstum pro Jahr	Big Data Business Modell Transformation Potential
Industrial	6	8	20-30%	Mittel
Mobility & Logistics	4	9	40-50%	Sehr hoch
Professional Services	5	8	25-35%	Hoch
Finance & Insurance	8	10	30-40%	Hoch
Healthcare	5	9	40-50%	Sehr hoch
Government/ Education	3	8	10-20%	Sehr hoch
Utilities	4	6	10-20%	Mittel
IT, Telco, Media	8	10	50-60%	Sehr hoch
Retail/Wholesale	2	7	20-30%	Sehr hoch

Tabelle 9: Transformationspotenzial durch Big Data; Quelle: Experton Group 2012



10 Einsatzbeispiele von Big Data in Wirtschaft und Verwaltung

Big Data wird bereits in vielen Unternehmen und Organisationen produktiv genutzt, wie die drei Dutzend Einsatzbeispiele überzeugend nachweisen. Der Mangel an Anwendungsbeispielen gilt als Hürde für den Erfolg am Markt – deswegen wurden für den Leitfaden auch einige anonymisierte Beispiele zugelassen, wenn sie interessante Einblicke eröffnen.

Die Beispiele sind geeignet, Manager anzuregen, in ihren Unternehmen Einsatzmöglichkeiten zu entdecken und Daten in Business Value zu verwandeln.

Hinweis: Das Kapitel 10 stellt am Markt existierende Anwendungen vor. BITKOM hat die rechtliche Zulässigkeit der Anwendungsbeispiele nicht überprüft und übernimmt hierfür keine Verantwortung. Es ist Aufgabe jedes Nutzers von Big Data, vor einem Einsatz zu prüfen, ob die Anwendung mit dem deutschen Datenschutzrecht und weiteren Rechtsvorschriften vereinbar ist.

Zur schnellen Orientierung sind die Einsatzbeispiele nummeriert (№01...№34) und

- Funktionsbereichen in Unternehmen und Organisationen sowie
- Wirtschaftszweigen

zugeordnet.

Wie bereits im Kapitel 7 werden sieben Funktionsbereichen betrachtet (vgl. Tabelle 10).

Kürzel	Funktionsbereich
MuV	Marketing und Vertrieb
FuE	Forschung und Entwicklung
PRO	Produktion
SuS	Service und Support
DuL	Distribution und Logistik
FRC	Finanz- und Risiko-Controlling
AOO	Administration, Organisation und Operations

Tabelle 10: Funktionsbereiche der Big-Data-Einsatzbeispiele

Im Kapitel 10 werden Big-Data-Anwendungen aus diesen sieben Funktionsbereichen vorgestellt. Darunter sind Beispiele aus neun Wirtschaftszweigen (vgl. Tabelle 11):

Kürzel	Wirtschaftzweig
ÖV	Öffentliche Verwaltung
HL	Handel und Logistik
AM	Automobilbau, Manufacturing
DI	Dienstleistungen
FD	Finanzdienstleistungen
MB	Maschinenbau
IW	Informationswirtschaft (Informationstechnik, Telekommunikation, Medien)
EE	Elektrotechnik und Elektronik
GW	Gesundheitswesen

Tabelle 11: Wirtschaftszweige der Big-Data-Einsatzbeispiele

		Wirtschaftszweige								
		ÖV	HL	AM	DI	FD	MB	IW	EE	GW
Funktionsbereich	MuV		№03 №05 №07 №10		№02 №06 №09 №12			№01 №04 №08 №11		
	FuE	№14			№13					№15
	PRO			№18			№16		№17	
	SuS			№19						№20
	DuL		№21							
	FRC					№22 №23 №24				
	AOO	№25 №31 №32	№27 №28 №30 №33		№34			№26 №29		

Die (mitunter nicht ganz eindeutige) Zuordnung der Big-Data-Einsatzbeispiele nach Funktionsbereichen und Wirtschaftszweigen zeigt die Tabelle 12.

Tabelle 12: Einsatzbeispiele für Big-Data - Übersicht nach Funktionsbereichen und Wirtschaftszweigen

Volumen
500
Millionen Datensätze
Analyse Abrufzahlen Web-Videos
(№01)

Volumen
100.000
I/O's pro Sekunde
Erhöhung des Transaktionsvolumens
(№02)

Geschwindigkeit
270
Millionen Preispunkte in zwei Stunden
Preisoptimierung im Einzelhandel
(№05)

Vielfalt
2,5
Milliarden Interaktionen im Monat – Reduktion der Analyse-Laufzeiten um Faktor 100 – polystrukturierte Datenquellen / Echtzeit-Analysen von »Social Intelligence« (№06)

Volumen
40%
Verbesserung
Prognose – 300 Millionen Datensätze wöchentlich / Verkaufsprognosen
(№07)

Volumen
50%
Kostenreduktion – 1 TB neue technische Daten wöchentlich
Vorhersage von Kundenkündigungen
(№08)

Geschwindigkeit
100 Millionen Keywords /
75 Millionen Domains Ranking von Websites, Such-Stichworten
(№09)

Geschwindigkeit
Liveanalysen früher in Stunden, jetzt in
Sekunden
Steuerung / Beschleunigung der Vertriebsprozesse (№10)

Geschwindigkeit
500
Millionen Verbindungen in Auswertung
Verminderung Fluktuation (№11)

Vielfalt 90% Zeiteinsparung bei Analysen Competitive Intelligence (Nº13)	Vielfalt 250.000 GPS-Daten pro Sekunde Echtzeit-Analyse für städtische Verkehrssteuerung (Nº14)	Vielfalt 3.5 GB pro Minute Überwachung/Diagnose komplexer technischer Systeme (Nº16)
Volumen 5 Mrd. Rohdatensätze in kürzester Zeit Optimierung von Geschäftsprozessen (Nº17)	Geschwindigkeit 4…60 Faktor der Geschwindigkeitsverbesserung Globale Planung und Steuerung (Nº18)	Volumen 10 TB Daten über die Wertschöpfungskette Ganzheitliche Qualitätsanalyse (Nº19)
Volumen 200 Mio. Posts im Internet pro Tag Analyse von Gesundheitsdaten (Nº20)	Geschwindigkeit 200.000 Input-Output-Operationen pro Sekunde / Management von Fahrzeugflotten (Nº21)	Vielfalt Millionen Transaktionen pro Tag Betrugsanalyse bei Kreditkarten (Nº22, Nº23)
Geschwindigkeit 8,8 Mrd. Value-at-Risk-Berechnungen in zwei Minuten / Berechnung Risiken im Portfolio einer Bank (Nº24)	Volumen >500 Mio. Bürger in 5 Jahren Identifikations-Service für Bürger Indiens (Nº25)	Volumen >1 Mrd. Ereignisse pro Monat – Verarbeitung in 3 Stunden / Versand vertrauenswürdiger E-Mails (Nº26)
Geschwindigkeit < 2s Retrieval eines Dokuments aus 5 Mrd. Dokumenten / Compliance im Paketversand über 10 Jahre (Nº27)	Volumen 220.000 Datenquellen werden überwacht Management eines komplexen IT-Systems (Nº28)	Volumen 240 Mrd. Datensätze in 18 h analysiert Komplexe Mustererkennung in Diagnoseplattform (Nº29)
Volumen 350 GB Daten in 24 h analysiert Monitoring IT-Infrastruktur (Nº30)	Vielfalt 13 TB mit 450 Millionen Einträgen – Retrieval in weniger als 0,3s Weltweite Fachrecherche in Patentdatenbanken (Nº31)	Volumen 25 Mrd. gefahrene km pro Jahr / Mustererkennung für Qualitätssicherung (Nº33)
Volumen 30 TB Datenvolumen Echtzeitsteuerung von Angeboten auf Social-Media-Plattform (Nº34)		

Abbildung 17: Ausgewählte Big-Data-Einsatzbeispiele in Kennziffern



10.1 Einsatzbeispiele aus Marketing und Vertrieb

10.1.1 (Nº01) Deutsche Welle – Nutzungsdaten und -analysen von Web-Videos auf einen Blick

Anwender	Deutsche Welle Kurt-Schumacher-Straße 3, 53113 Bonn www.dw.de
Ansprechpartner	Werner Neven, Markt- und Medienforschung, Tel.: +49 (0)228 429 3104 werner.neven@dw.de
Anbieter	The unbelievable Machine Company GmbH Grolmanstr. 40, 10623 Berlin www.unbelievable-machine.com
Ansprechpartner	Klaas Bollhoefer, Data Scientist Tel.: +49 (0)30 889 2656 22 klaas.bollhoefer@unbelievable-machine.com
Problem	Die Deutsche Welle (DW) produziert täglich eine Vielzahl an Web-Videos und -Formaten für seine digitalen Angebote, unter anderem für das eigene Media Center auf http://mediacenter.dw.de. Darüber hinaus werden die Videos über ein zentrales Media-Asset-Management-System an alle wichtigen Web-Video-Plattformen ausgespielt, und dies aufgrund der internationalen Ausrichtung der DW derzeit in mehr als 10 Sprachen. Neben derzeit 16 YouTube-Kanälen der DW werden auch Kanäle auf MyVideo, Sevenload und Dailymotion mit den Videos bespielt, weitere – vorwiegend asiatische – Plattformen sind in Planung. Die Markt- und Medienforschung der Deutschen Welle möchte die Nutzungszahlen ihrer Videos auf allen Plattformen monitoren und einen näheren Einblick in das Nutzungsverhalten bekommen. Dabei werden nicht nur die eigenen Kanäle, sondern auch die anderer internationaler Sender betrachtet. Ein einheitlicher Zugang zu den Daten auf den verschiedenen Plattformen sowie die Möglichkeit übergreifender, systematischer Auswertungen sind von zentraler Bedeutung.
Lösung	Hohe Anforderungen in Bezug auf Datenmenge, hohe Verarbeitungsgeschwindigkeit und unterschiedliche Datenformate sprachen eine klare Sprache und führten zur Implementierung einer Big-Data-Lösung auf Basis von Hadoop und HBase. Die Aufgabe für The unbelievable Machine Company lautete, ein System zu konzipieren und zu implementieren, das in der Lage ist, die Abruf- und Nutzungszahlen aller Videos von den unterschiedlichen Plattformen zu importieren, zu normalisieren, in ein einheitliches Format zu überführen und nicht zuletzt in einem für die Zwecke der Mitarbeiter der DW optimierten, browserbasierten Bedieninterface darzustellen. Es galt, eine Big-Data-Lösung zu entwickeln und in Betrieb zu nehmen, die auf der einen Seite große Datenmengen importieren und parallel verarbeiten kann und diese auf Dauer persistent in einer Datenbank vorhält. Auf der anderen Seite sollten auf der Analyse-Ebene hochperformante Datenverarbeitungs- und intelligente Analyse-Module als Komponenten implementiert werden. So wurde beispielsweise auch ein Machine-Learning Classifyer entwickelt, der automatisiert eine Gruppen- und Sendungsformatzuordnung vornimmt.

Big-Data-Merkmale	Infrastruktur: Die Big-Data-Lösung der DW besteht aus einem dedizierten Hadoop-Cluster. Als Datenbank kommt HBase zum Einsatz. Die Abruf- und Nutzungszahlen werden stündlich von den Web-Video-Plattformen abgegriffen, pre-processed und in HBase importiert. Regelmäßig laufen mehrere Map/Reduce-Jobs in Hadoop, um übergreifende Auswertungen zu erstellen. Das interaktive, tabellenbasierte Auswertungs- und Analyse-Interface für die Mitarbeiter der Markt-Medienforschung der DW wurde in BIRT implementiert. Der hochperformante und sichere Betrieb erfolgt nun seit mehr als 1,5 Jahren im mehrfach zertifizierten *um DataCenter in Berlin. Merkmale Bis heute wurden mehr als ■ 500 Millionen Datensätze für ■ mehr als 50.000 Videos importiert (Stand: Ende August 2012), aggregiert und analysiert.
Nutzen	Der primäre Nutzen für die Mitarbeiter der Deutschen Welle liegt darin, dass sie ■ EIN Interface haben, das ihnen ■ ALLE Abruf- und Nutzungszahlen ■ ALLER Videos ■ auf EINEN Blick liefert. Die Markt- und Medienforschung wird in die Lage versetzt, all diese Daten über ein einheitliches Web-Interface zu betrachten, zu analysieren und so in ihre Arbeitsergebnisse einfließen zu lassen. Dadurch sparen sie nicht zuletzt auch viel Zeit und Mühen. Der Blick auf die Web-Video-Nutzung anderer europäischer Sender erlaubt Vergleiche auf inhaltlicher und struktureller Ebene. Die seit Beginn des Projekts aggregierten und einheitlich strukturierten Daten können in Zukunft auch für tiefergehende Analysen genutzt werden, beispielsweise das Auffinden von Abrufmustern und daraus resultierend für Prognosen.
Lessons learnt	Big Data Infrastruktur: ■ Hadoop/HBase sind ausgereifte und in Summe sehr robuste Technologien, nichtsdestotrotz ist ein hohes Maß an Erfahrung und Expertise in den Bereichen Implementierung und Operating nötig, um diese Big-Data-Technologien professionell in den Regelbetrieb zu nehmen ■ HBase skaliert auch über große Datenmengen exzellent (wenn richtig konfiguriert und in der Infrastruktur verortet). Es ist nicht zu erwarten, dass das Projekt in den nächsten Jahren an Grenzen stößt. Big Data Use Case: ■ Essentiell auch für Big-Data-Projekte sind eine klare Aufgabenstellung, Fokus auf die Lösung und die Nutzer dieser Lösung (weniger auf neueste IT-Technologien) und nicht zuletzt auch ein Gespür für Usability und Funktionsumfang eines Reporting- / Analyse-Dashboards. Weniger ist hier meistens mehr. ■ An dieser Stelle kommen die Funktion und die Aufgaben eines »Data Scientist« zum Tragen, der ein Big-Data-Projekt von der Formulierung der Aufgabenstellung über die Technologie bis zum Nutzerinterface (»machine- or human-readable«) begleitet und die Gesamtkonzeption der Use- und Business-Cases vornimmt.

10.1.2 (Nº02) DeutschlandCard GmbH – Effiziente IT-Infrastruktur für Big Data

Anwender	arvato Systems / DeutschlandCard Migration und Betrieb der Datenbank-Infrastruktur der DeutschlandCard GmbH auf einer Oracle Exadata-Lösung. Anbieter arvato Systems GmbH An der Autobahn 200, 33333 Gütersloh info@arvato-systems.de www.arvato-systems.de Deutschland Card
Problem	Die DeutschlandCard GmbH hat 2011 zwei neue Partner gewonnen, die zu einer erheblichen und stetigen Steigerung des täglich zu verarbeitenden Transaktionsvolumens sowie zu einer kontinuierlichen Steigerung des Datenvolumens im DWH führten. Neben der Verarbeitung dieser Datenmengen sollte zusätzlich das Zeitfenster für die ETL-Prozesse deutlich optimiert und Raum für zusätzliche Funktionalitäten und Sonder-Reports geschaffen werden. Die Grenze der klassischen Architektur mit 4 GB/s Read-Durchsatz und in Spitzenzeiten bis zu 100.000 I/O's pro Sekunde erreicht. Aus Business-Sicht kam keine Begrenzung des Datenbestandes infrage. Der Bedarf an einer performanteren DWH-Umgebung mit hohem Storage-Volumen stand damit im Vordergrund, verbunden mit der Forderung nach geringstmöglichem Migrationsrisiko und kurzer Downtime. Größere Änderungen an der komplexen Applikations- und Datenbanklandschaft der DeutschlandCard sollten möglichst vermieden werden.
Lösung	Die Migration der Datenbank-Infrastruktur für die DeutschlandCard GmbH erfolgte von einem 4-Knoten Oracle-RAC auf eine Exadata X2-2 Database Machine, nachdem in einem Proof of Concept die geforderten Performanceverbesserungen nachgewiesen wurden. Die Migration erfolgte teils mit klassischem Import/Export-Verfahren, teils über Metadatenanlage mit Modifikationen, die die Nutzung der neuen Performance-Features der Exadata Veränderungen im Datenmodell nötig machte. Die eigentliche Migration der hoch integrierten Applikationslandschaft war komplex und mit hohen Anforderungen an die Sicherheit mit 2 kompletten Tests und einer Generalprobe verbunden. Dabei durfte die Downtime von max. 4 Stunden für die Online-Applikation nicht überschritten werden. Im IT-System-Management hat die arvato Systems GmbH die bestehenden Schnittstellen und Betriebsprozesse im Sinne einer ganzheitlichen Betrachtung weiterentwickelt. Zusammen mit Spezialisten der arvato IT Services, Oracle und Trivadis wurde die optimale Kombination der zahlreichen Möglichkeiten (Smart Scan Verfahren, HCC Compression) der Exadata-Technologie ermittelt und umgesetzt.
Big-Data-Merkmale	**Volume:** Das Datenwachstum liegt im monatlichen Mittel in einem dreistelligen GB-Bereich. Das gespeicherte Datenvolumen liegt im zweistelligen TB-Bereich. **Variety:** Bei den aktuellen Daten handelt es sich um strukturierte Informationen; es ergeben sich kontinuierlich neue Anforderungen an zu speichernde Daten, weitere Partner, mobile und unstrukturierte Datenbestände. In Abhängigkeit von der Art der Daten werden diese optimal gespeichert und verwaltet. **Velocity:** Das DWH steht morgens für die Business Intelligence Experten der DeutschlandCard mit den Vortagesdaten bereit. Mit der neuen Architektur sind auch komplexe Ad-hoc-Auswertungen realisierbar und i.d.R. im Minutenbereich ausführbar.

Nutzen	Die tägliche Aktualisierung des DWH wurde signifikant verkürzt, sodass Zeit für neue Auswertungen bleibt, die der DeutschlandCard eine Geschäftserweiterung ermöglichen. Die OLTP-Prozesse (Online Transaction Processing) sind deutlich schneller, die Laufzeit der OLAP-Prozesse (Online Analytical Processing) wurde signifikant reduziert. Einzelne komplexe Read/Write-Operationen verbesserten sich ebenfalls signifikant. Ergebnisse für komplexe Abfragen werden jetzt hochperformant geliefert, da die Verarbeitungslast auf alle Storage-Knoten verteilt wird. Diese Skaleneffekte ermöglichen der DeutschlandCard einen weiteren Geschäftsausbau. Der Datenbankserver als bisheriger Engpass ist deutlich entlastet. Zusätzlich ist jetzt eine Parallelität zwischen ETL-Prozessen und Standard-Report-Generierung ohne gegenseitige Beeinträchtigung möglich. Eine redundant ausgelegte Konsolidierungsplattform für mehrere Datenbanken steht bereit.
Lessons learnt	Eine ganzheitliche Betrachtung einer solchen Big-Data-Lösung ist notwendig, um die Performanceanforderungen der Applikation mit den Kosten und dem Business-Nutzen in Einklang zu bringen. Nur ein minutiöser Migrationsplan mit mindestens einer kompletten Generalprobe inklusive Fallback-Test sichert die Betriebssicherheit einer solch komplexen Applikation mit ihren zahlreichen Schnittstellen zu externen Partnern. Die sehr gute und direkte Teamarbeit zwischen arvato Systems, arvato IT Services, Oracle und Trivadis war ein wesentlicher Erfolgsfaktor.

10.1.3 (Nº03) dm – Mitarbeitereinsatzplanung

Anwender	Roman Melcher, Geschäftsführer IT, dm
Anbieter	Blue Yonder GmbH & Co. KG
Ansprechpartner	Dunja Riehemann dunja.riehemann@blue-yonder.com
Problem	In der Vergangenheit haben die Filialverantwortlichen die Mitarbeiterkapazitäten auf Basis einfacher Hochrechnungen sowie ihrer Erfahrungswerte geplant und in das System eingegeben. Danach errechnete das Programm den Mitarbeiterbedarf pro Tag. Normalerweise hat das gut funktioniert, allerdings geriet das Verfahren in Sondersituationen an seine Grenzen. Die Folge waren Über- oder Unterbesetzungen.
Lösung	dm führte für die Vorhersage der Tagesumsätze die Predictive-Analytics-Suite von Blue Yonder ein. Vier bis acht Wochen im Voraus tragen sich die Mitarbeiter der jeweiligen Filiale nach ihren persönlichen Präferenzen in den Bedarfsplan ein. Seit dem Blue Yonder-Projekt können sie sich auf die einmal abgestimmte Planung verlassen. Kurzfristige Änderungen sind selten geworden.
Big-Data-Merkmale	Neben den Tagesumsätzen fließen auch die Paletten-Anliefer-Prognosen der Verteilzentren und filialindividuell einstellbare Parameter wie die Öffnungszeiten in die Planung ein. Beides wird benötigt, um den Mitarbeiterbedarf möglichst genau zu ermitteln. So wirkt sich etwa der Wareneingang erheblich auf den Personalbedarf einer Filiale aus. Auch die Frage, in welcher Zeit der prognostizierte Umsatz getätigt werden soll, ist für die Kapazitätsplanung wesentlich. Durch die Einspeisung der NeuroBayes®-Prognosen lassen sich der Kundenandrang und damit das tagesbezogene Umsatzvolumen sehr viel genauer abschätzen. Darüber hinaus werden alle verlässlichen Daten einbezogen, auch externe. Das können Markttage, Ferien im Nachbarland oder eine Baustelle an der Zufahrtsstraße sein. Auch die Wettervorhersage lässt sich künftig berücksichtigen. Derzeit generiert die Lösung jede Woche 225.000 Prognosen für alle Filialen – den geplanten täglichen Filialumsatz für die nächsten Wochen. Ab Mitte Oktober 2012 werden es 450.000 Prognosen jede Woche sein.
Nutzen	Auf diese Weise können folgende Vorteile realisiert werden: ■ Ermittlung der Umsätze pro Filiale auf Tagesebene ■ Verlässliche Prognose für die Mitarbeitereinsatzplanung ■ Verbesserung der Mitarbeiter- und Kundenzufriedenheit.

dm
HIER BIN ICH MENSCH
HIER KAUF ICH EIN

10.1.4 (Nº04) etracker – Verbindung konventioneller IT-Systeme mit Big-Data-Technologien

Anwender	Die etracker GmbH aus Hamburg ist mit mehr als 110.000 Kunden ein in Europa führender Anbieter von Produkten und Dienstleistungen zur Optimierung von Websites und Online-Marketing-Kampagnen. etracker wurde mehrfach mit Innovationspreisen ausgezeichnet. Unter anderem nutzen Henkel, HSBC Trinkaus & Burkhardt, Lufthansa Worldshop, Spiegel Verlag und T-Online die Produkte.
Anbieter	ParStream GmbH, Große Sandkaul 2, 50667 Köln, www.parstream.com, +49 221 97761480; Joerg.Bienert@Parstream.com
Problem	Bei Analyse-Systemen, die auf herkömmlichen Datenbanken basieren, müssen Reports vorab definiert werden. Jede Änderung oder Ergänzung führt zu langen und aufwändigen Neuberechnungen. Dagegen hängt der Erfolg einer Website maßgeblich davon ab, das Verhalten der Nutzer flexibel analysieren zu können. So unterschiedlich Websites sind, so unterschiedlich sind auch Website-Betreiber in ihren Wünschen an ein Analysetool. Vordefinierte Reports treffen damit nur bis zu einem bestimmten Punkt den Bedarf des einzelnen Website-Betreibers und er vermisst die Möglichkeit, während seiner Analyse neue Abfragen zu definieren. etracker wollte ein Tool entwickeln, mit dem Anwender zum einen Massendaten ohne lange Wartezeiten auswerten und zum anderen den Fokus bei Datenanalysen flexibel und dynamisch wechseln können..
Lösung	etracker hat über mehrere Jahre eine neuartige, zweistufige Datenbank-Technologie entwickelt. Zunächst werden sämtliche erfasste Daten als Rohdaten gespeichert. Im zweiten Schritt kommt die ParStream-Datenbank zum Einsatz. Durch diese Verknüpfung entstand das neue Tool, etracker Dynamic Discovery, welches Anwendern ermöglicht, die Perspektive auf die Daten jederzeit per Drag & Drop zu verändern und weitere Kennzahlen ganz flexibel in die Auswertung mit aufzunehmen oder wieder herauszufiltern. Der Anwender kann aus einem für das jeweilige Business-Modell relevanten Set an Parametern beliebig auswählen und diese für seine Analysen individuell kombinieren. Das Ergebnis erhält der Anwender auch bei komplexen Anfragen innerhalb von wenigen Sekunden.
Big-Data-Merkmale	Die Herausforderung seitens etracker ist ein typisches Big-Data-Problem: Es werden sehr große Datenmengen über lange Zeiträume gespeichert, die mit großer Dynamik wachsen. Diese Daten sollen in Echtzeit gefiltert werden, bei hunderten von gleichzeitigen Nutzern. Solche Anforderungen lassen sich mit konventionellen Datenbanktechnologien nicht mehr bewältigen.
Nutzen	Website-Betreiber profitieren mit etracker Dynamic Discovery von sehr schnellen, dynamischen Auswertungen ihrer Massendaten. Sie können dementsprechend zeitnah reagieren und verschaffen sich so einen Wettbewerbsvorteil am Markt. Eine SEA-Analyse lässt sich beispielsweise in Hinblick auf Suchmaschinen, Suchwörter oder den tatsächlich erzielten Umsatz durchführen. Dabei können unterschiedliche Aspekte beliebig per Drill-down-Funktion mit einbezogen werden: Einstiegsseiten, gesehene und bestellte Produkte bzw. Kategorien oder auch im Hinblick auf Bouncer, stehen gelassene Warenkörbe oder Stornoraten, etc. Mit relationalen Datenbanken wäre eine solche Dynamik bei diesen Datenmengen nicht oder nur unter erheblich höheren Investitionen möglich gewesen. Der Einsatz der ParStream Big-Data-Analytics-Technologie ist somit eine der wesentlichen Grundlagen für die wirtschaftliche Umsetzung eines ganz neuen, innovativen Leistungsangebots.
Lessons learnt	Die Kombination aus herkömmlicher Datenbanktechnologie mit der spezialisierten ParStream-Software ermöglicht besonders effiziente Lösungen und verbessert neben dem Funktionsumfang auch die Wirtschaftlichkeit deutlich.

10.1.5 (№05) Macy's – Preisoptimierung

Anwender	Macy's ist ein amerikanisches Handelsunternehmen (stationär und online)
Anbieter	SAS Institute
Problem	Macy's gehört zu den größten überregional tätigen Händlern in den USA. In den 800 Filialen wird ein großes, mehrere zehntausend Artikel umfassendes Sortiment angeboten. Um möglichst optimale Preise anzubieten, geht das Unternehmen auf die jeweils standortspezifischen Unterschiede der einzelnen Filialen ein. Wenn für eine bestimmte Produktklasse ein starker Wettbewerber benachbart ist, werden die Preise in dieser Filiale aggressiver nach unten angepasst, um auf jeden Fall wettbewerbsfähig zu sein. Ist kein Wettbewerber vorhanden, ist diese Notwendigkeit nicht gegeben. So ergeben sich über das gesamte Sortiment und alle Standorte etwa 270 Millionen Preispunkte. Auf der Basis der in der Vergangenheit erfolgten Abverkäufe bestimmter Waren (etwa zwei Terabyte an Daten) wurden bisher wöchentlich neue Preise für die Sortimente berechnet – was ungefähr 30 Stunden Rechenzeit in Anspruch nahm. Da Macy's sieben Tage pro Woche geöffnet hat, konnten regelmäßig bestimmte Abverkäufe gar nicht in die Analyse aufgenommen werden. Man behalf sich mit Teilsortiments-Optimierungen.
Lösung	Durch die Umstellung der vorhandenen Infrastruktur auf optimierte Datenhaltung und dem Einsatz von In-Memory-Technologie war es möglich, die Analyse über das gesamte Sortiment auf eine Zeit unter zwei Stunden zu drücken
Big-Data-Merkmale	Das große Sortiment mit entsprechend vielen Daten zum Abverkauf hat dazu geführt, dass das Datenvolumen immer weiter gewachsen ist. Die neue Lösung kann nun skalieren und liefert gleichzeitig die Geschwindigkeit, Preise im Tagesverlauf mehrfach anpassen zu können – trotz Datenvolumina von mehr als zwei Terabyte pro Analyse.
Nutzen	Macy's ist damit in der Lage mehrfach pro Tag neue Preise stellen zu können und im gesamten Sortiment noch besser auf die lokalen Wettbewerber reagieren zu können – und das auf einer analytisch fundierten Basis, die allein es ermöglicht, auf breiter Front eine optimierte Preispolitik zu betreiben. Big Data Analytics optimiert hier an einem Kernprozess der Branche – der Preisfindung. Gerade durch das hohe Datenvolumen gibt es gerade im Handel ein hohes Potential, signifikante Wettbewerbsvorteile zu erzielen.[65]
Lessons learnt	Der Businessnutzen zeigt sich erst, wenn Prozesse, die aufgrund fehlender Möglichkeiten bewusst eingeschränkt wurden, geändert werden. In diesem Fall ist es die früher gar nicht mögliche sehr viel häufigere Preisoptimierung auf dem Gesamtsortiment. Auch können nun sehr viel aktuellere Abverkaufszahlen mit in die Analyse einbezogen werden.

[65] Vgl. Lünendonk®-Trendpapier 2012 »Big Data im Handel – Chancen und Herausforderungen« vom 16.07.2012

10.1.6 (№06) MZinga – Echtzeit-Analysen von »Social Intelligence«

Anwender
Navdeep Alam, Director of Data Architecture, Mzinga

Anbieter
Teradata Aster

Ansprechpartner
Dr. Andreas Ribbrock, andreas.ribbrock@teradata.com

Problem
Bei dem Altsystem waren die Latenzzeiten bei Anfragen zu lang, um die Erkenntnisse aus Big-Data-Analysen zeitnah in Geschäftsprozessen zu verwenden. Außerdem ließen sich nicht alle Analysen effizient mit SQL allein umsetzen – z. B. Analysen auf sozialen Graphen.
An das neue Big-Data-System wurden folgende Anforderungen gestellt:
- Analysen aus dem Bereich »Social Intelligence« müssen in nahezu Echtzeit möglich sein
- Die Analysen sollten auf der gleichen Plattform stattfinden, die auch zur Speicherung der ca. 2.5 Milliarden Interaktionen im Monat genutzt werden wird.
- Durch den Wegfall von ETL-Strecken und »Embedded Analytics« sollen die Scoring-Modelle schneller auf die Daten angewendet werden.
- Die Plattform soll skalierbar sein, um dem wachsenden Datenvolumen über die Zeit gerecht werden zu können.

Lösung
Die Lösung besteht in der Parallelisierung von komplexen analytischen Problemen auf einer massiv-parallelen Plattform unter der Verwendung vom Hersteller vorgefertigter Algorithmen, um bei Parametern wie Umsetzungszeit und -güte den dynamischen Anforderungen im Bereich Big-Data-Analysen gerecht zu werden.
Die Kombination von SQL und MapReduce ermöglicht komplexe Analysen auf großen sozialen Graphen, um Kundenverhalten schneller und besser zu verstehen. Aufgrund des SQL-Interfaces können Mehrwerte aus Analysen in kurzer Zeit auf vielen Terabyte an Daten mit den bestehenden Tools und Kenntnissen erreicht werden.
Es werden folgende Verfahren eingesetzt: Zeitreihenanalysen, Musteranalysen, Analysen auf sozialen Graphen.

Big-Data-Merkmale
Volume:
Analysen werden auf einem Datenbestand im Terabyte-Bereich durchgeführt, der täglich um Millionen von Interaktionen in sozialen Netzwerken wächst.
Variety:
Daten werden aus unterschiedlichsten poly-strukturierten Datenquellen zusammengeführt und auf innovative Weise analysiert.
Velocity:
Neue Fragestellungen werden in kurzer Zeit durch neue Datenanalysen beantwortet. Ergebnisse werden in nahezu Echtzeit bereitgestellt.

Nutzen
Der Nutzen lässt sich an drei Merkmalen aufzeigen:
- Relevanz: Geschäftsprozesse können in nahezu Echtzeit mit den benötigten Daten versorgt und neue Einsichten kurzfristig gewonnen werden.
- Performanz: Die Reduktion von Laufzeiten von vielen Stunden auf wenige Minuten ermöglicht mehr und komplexere Analysen auf großen Datenmengen.
- Lineare Skalierung: Die massiv parallele Architektur ermöglicht die Bereitstellung von neuer Hardware entlang des Wachstums der Datenmenge bzw. Anzahl der Nutzer.

Lessons learnt
»Analysen sind bei uns nun nicht mehr wenigen Leuten mit Spezialkenntnissen zugänglich, sondern zu Echtzeit-Anwendungen für viele geworden.«

10.1.7 (№07) Otto – Verbesserung der Absatzprognose

Anwender	Michael Sinn, Direktor Angebots- und Category Management Support, Otto
Anbieter	Blue Yonder GmbH & Co. KG
Ansprechpartner	Dunja Riehemann, dunja.riehemann@blue-yonder.com
Problem	Als »Data-Driven Company« setzt OTTO schon immer auf die Auswertung umfangreicher Datenmengen zur Entscheidungsunterstützung. Ein eigener Unternehmensbereich »Business Intelligence« beschäftigt sich bei OTTO damit. Allerdings stießen die klassischen Prognosemethoden in der komplexen Online-Welt schnell an ihre Grenzen. Die Herausforderung war: ■ Die Lieferbereitschaft sollte deutlich erhöht und damit verbundene Out-of-stock Situationen reduziert werden. ■ Das Bestandsmanagement galt es zu optimieren und Überhänge zu vermeiden. ■ Die technische Herausforderung war, komplexe Zusammenhänge in Massendaten zu erkennen und präzise Prognosen daraus zu erstellen.
Lösung	Die Predictive-Analytics-Software von Blue Yonder ist eine Kombination aus neuronalen Netzen und statistischen Methoden. Nach einer umfangreichen Untersuchung von Otto war das selbstlernende System anderen Verfahren deutlich überlegen. Denn NeuroBayes® erkennt relevante Zusammenhänge in Massendaten und liefert mit seiner robusten Methodik äußerst präzise Prognosen. Diese unterstützen den Einkauf dabei, die Stückzahlen bei den Lieferanten so nah wie möglich am Ist zu platzieren und so letztendlich den Gewinn zu steigern.
Big-Data-Merkmale	Jährlich werden mehr als 1 Milliarde Einzelprognosen erstellt, die anhand einer Vielzahl von Faktoren berechnet werden. Dafür ist ein souveräner Umgang mit Massendaten erforderlich: Täglich bzw. wöchentlich fließen bis zu 135 Gigabyte oder 300 Millionen Datensätze ins System. Für die Prognose spielen dabei Faktoren wie der Bewerbungsgrad eines Artikels online und offline sowie spezifische Artikeleigenschaften und Umfeldbedingungen eine entscheidende Rolle.
Nutzen	Der Vorteil ist eine deutlich bessere Lieferbereitschaft für den Kunden und eine höhere Wirtschaftlichkeit. ■ Prognoseverbesserungen – je nach Angebotsartikel – bis zu 40% ■ Steuerung der Kurz-, Mittel- und Langfristprognosen von Otto durch lernende Systeme ■ Auswertung riesiger Datenmengen in Echtzeit.
Lessons learnt	»Wir haben erkannt, dass für unsere Anforderungen ein selbstlernendes System notwendig ist, das sich stetig ändernde Einflussfaktoren wie Ansprache und Artikel-Ranking oder im Printbereich Seitenanteil und Katalogausstoßmenge berücksichtigt. Damit steigt unsere Prognosequalität kontinuierlich und die prognostizierten Absatzmengen werden immer präziser. Außerdem können wir uns frühzeitig auf künftige Entwicklungen einstellen.«.

10.1.8 (№08) Satelliten-TV Anbieter – Customer Churn und »Pay-per-View«-Werbeoptimierung (Pilot)

Anwender	Satelliten-TV-Anbieter
Anbieter	Wipro Technologies, Analytics & Information Management, Mario Palmer-Huke (mario.palmerhuke@wipro.com)
Problem	Der Kunde möchte basierend auf Zustands- und Fehlerdaten der Set-Top Box (STB) ein statistisches Modell für die Vorhersage von Kundenkündigungen (»churn«) entwickeln. Die korrekte Funktion der STB wurde als ein entscheidendes Kriterium für die Kundenzufriedenheit identifiziert. Darüber hinaus sollen die Daten der STB genutzt werden, um die Werbeschaltungen für »Pay Per View« (PPV) zu optimieren.
Lösung	Gemeinsam mit dem Kunden wurden auf Basis beispielhafter STB-Daten der letzten Jahre Muster- und Vorhersagemodelle für das potentielle Kündigungsverhalten entwickelt. Darüber hinaus wurden die STB-Daten in einer neu installierten Appliance mit anderen Daten verknüpft, um das optimale Placement der PPV-Werbung zu berechnen. Im Rahmen eines Piloten wurden verschiedene Appliance-und Auswerteplattformen evaluiert, um die Lösung mit der besten Kosteneffizienz (TCO) für den Kunden zu identifizieren.
Big-Data-Merkmale	Für die initiale Analyse standen 80 Mrd. Datensätze der letzten Jahre zur Verfügung. Darüber hinaus generieren die STB ca. 1 TB neue technische Daten pro Woche. Weitere denkbare Anwendungen (Echtzeit-Programmempfehlungen) erfordern darüber hinaus fast eine Echtzeitverfügbarkeit der STB-Daten.
Nutzen	Die neue Architektur erlaubt die Einführung komplexerer Modelle zur Vorhersage des Kundenkündigungs-Verhaltens zu ca. 50% der Kosten des bestehenden, traditionellen Enterprise Data Warehouses. Darüber hinaus konnte anhand der Beispieldaten gezeigt werden, dass durch die Werbeoptimierung der Umsatz des PPV um ca. 30% gesteigert werden kann. Neben der Produktivsetzung des Piloten werden im Moment weitere Anwendungen[66] diskutiert.
Lessons learnt	Im Rahmen des Piloten wurden nicht nur die technischen Komponenten auf ihre Sinnhaftigkeit und Funktionalität getestet, sondern auch mittels eines Datensamples der Nutzen des Anwendungsfalles überprüft. Ein positives Ergebnis macht eine folgende Investitionsentscheidung deutlich einfacher, da mittels der PoC-Ergebnisse der kommerzielle Nutzen leichter vorhergesagt werden kann.

66 z. B. Echtzeit-Programmempfehlungen, individualisierte Paketangebote basierend auf den Sehgewohnheiten

10.1.9 (Nº09) Searchmetrics – Realtime-Abfragen und -Auswertungen auf Milliarden von Datensätzen

Anwender

Searchmetrics ist einer der international führenden Hersteller von SEO Analytics Software (SEO = engl. für Suchmaschinen-Optimierung).
Die Searchmetrics GmbH ist Pionier und international führender Anbieter von Search- und Social Analytics Software. Mit einer einzigartigen Serverinfrastruktur und Softwarelösungen wie den SEO Tools werden sehr große Datenmengen über das Ranking von Websites, Such-Stichworten, Backlinks und jeweils relevanten Wettbewerbergruppen und aggregiert und ausgewertet. Basis des Erfolgs sind die selbst entwickelten und weltweit einzigartigen Software-as-a-Service-Lösungen, Searchmetrics Suite und Essentials. Damit wird – analog zu den bereits weit verbreiteten Methoden und Techniken der Web Analytics – nun auch Search Analytics möglich. Denn die Software Lösungen erlauben es erstmalig, alle Aktivitäten im Bereich Suchmaschinen-Optimierung strukturiert zu analysieren, mit denen der Wettbewerber zu vergleichen und zu optimieren. Und ist damit in Sachen Datenmenge, Qualität und Geschwindigkeit herkömmlichen Tools weit überlegen. Namhafte Kunden aus Deutschland wie bild.de, ZEIT, Lufthansa, ProSieben, ImmoWelt, BASE oder T-Online haben Searchmetrics bereits ihr Vertrauen geschenkt.



Anbieter

ParStream GmbH, Große Sandkaul 2, 50667 Köln
www.parstream.com, Tel.: +49 221 97761480
Joerg.Bienert@Parstream.com

Problem

Searchmetrics beobachtet weltweit die führenden Suchmaschinen und zeichnet regelmäßig Organic und Paid Search Rankings und andere relevante Daten für über 100 Millionen Keywords und über 75 Millionen Domains auf. Daraus werden enorm große Datenpools gebildet, die teilweise mehrere Jahre zurückreichen. Für die Berechnung der Ergebnisse müssen diese großen Datenmengen zunächst umfassend bearbeitet und analysiert werden. Die zweite Herausforderung ist die Bereitstellung einer Online-Analysefunktionalität für den gesamten Datenbestand. Die bis zum Sommer 2011 eingesetzte Lösung auf Basis relationaler Datenbanken wie MySQL oder Oracle konnte die große Datenmenge nicht mehr in ausreichender Geschwindigkeit verarbeiteten.

Lösung

Auf den o.a. Datenpools mit Keyworddaten verschiedener Länder – bei denen über 7 Terabyte an Daten importiert werden müssen – verwendet Searchmetrics ParStream, um Realtime-Abfragen und Auswertungen auf über 10 Milliarden Datensätzen zu realisieren, mit denen Anwender die Online-Marketing-Aktivitäten auch von Wettbewerbern detailliert analysieren können.

Big-Data-Merkmale

Die Herausforderung bei Searchmetrics ist ein klassisches Big-Data-Problem, das mit herkömmlicher Datenbanktechnologie nicht mehr bewältigt werden konnte. Durch den Einsatz der Big-Data-Analytics-Plattform ParStream konnte diese Grenzen überwunden sowie neue Services bereitgestellt und Kosten gesenkt werden.

Nutzen

Mit der in der Vergangenheit eingesetzten Lösung war ein weiteres Datenwachstum ohne größere Investitionen nicht mehr möglich. Der Einsatz von ParStream war somit ein Bestandteil für weiteres Wachstum und eine internationale Expansion des Geschäfts sowie die Bereitstellung neuer Funktionalitäten.

Lessons learnt

Neue Daten können den Kunden auch im großen Umfang schneller zur Verfügung gestellt und neue Funktionalitäten bereitgestellt werden. Gleichzeitig konnte die Anzahl der erforderlichen Server gesenkt werden.

10.1.10 (№10) Schukat Electronic – Live-Analyse von Auftragsdurchlaufzeiten im Dashboard

SCHUKAT electronic

Anwender
Georg Schukat, Geschäftsführer, Schukat Electronic

Anbieter
SAP AG

Ansprechpartner
Boris Michaelis
SAP Deutschland AG & Co. KG
boris.andreas.michaelis@sap.com

Problem
Auch Mittelständler müssen die extrem wachsenden Datenmengen managen. Die bestmögliche Steuerung der geschäftskritischen Prozesse ist notwendig, um in der zunehmenden Dynamik des Wettbewerbs bestehen zu können. 1964 wurde Schukat Electronics in Monheim am Rhein gegründet und ist bis heute als unabhängiges Unternehmen in Familienbesitz. Mit 200 Herstellern im aktiven, passiven und elektromechanischen Bereich verbindet Schukat Electronics eine langjährige Partnerschaft. Den Kunden bietet das Unternehmen als Franchisepartner Entwicklungs- und Logistiksupport. Darüber hinaus ist Schukat ein Katalogdistributor für 9000 B2B-Kunden in 50 Ländern. Auswertungen über die verschiedenen Systeme im Vertrieb konnten nur mit erheblichem IT-Aufwand und mit entsprechendem Zeitverzug erstellt werden.

Lösung
Schukat Elektronic setzt nun SAP-HANA-In-Memory-Technologie und SAP-Analytics-Lösungen produktiv ein. Prozesse und Laufzeiten im Vertrieb werden optimiert auf Basis verschiedenster Systeme, z. B. Aufträge, Transportverfolgung etc. Alle Daten sind somit in einem System verfügbar und können nun von den Anwendern online ausgewertet werden.

Big-Data-Merkmale
Zusammenführung verschiedenster großer Datenquellen aus ERP- und Vertriebs-Systemen sowie Webshop und Versandsystem in einer Datenbank für Liveanalysen. Damit konnte die Geschwindigkeit (Velocity) von Stunden auf Sekunden reduziert werden.

Nutzen
- Realtime-Daten zur Steuerung bzw. Beschleunigung der Vertriebsprozesse
- Erhöhung der Servicequalität
- Gesteigerte Kundenzufriedenheit
- Reduzierung Aufwände in der IT, z. B. für Datenbankspezialisten.

Lessons learnt
Big Data Analytics ist nicht nur eine Herausforderung für Großunternehmen. Auch der Mittelstand muss sich immer mehr mit diesem Thema beschäftigen, um im internationalen Wettbewerb erfolgreich zu sein. Das Anwendungsbeispiel verdeutlicht den Nutzen im Vertrieb, aber auch z. B. in der Produktion mit Sensordaten etc. gibt es vielfältige Szenarien in den Fachabteilungen.

10.1.11 (Nº11) Telecom Italia – Minimierung der Kundenfluktuation

Anwender	Telecom Italia ist ein Festnetz- und Mobilfunk-Provider.
Anbieter	Hewlett-Packard GmbH, HP IM&A, Wulf Maier
Problem	Das Unternehmen bietet Telekommunikationstarife in einem hoch kompetitiven Markt an, in dem Telefonnutzer ihre mobilen Handytarife immer schneller wechseln können und diese so auf ihre persönlichen Telefongewohnheiten anpassen. Telecom Italia hat den Anspruch, diese Kundenfluktuation zu minimieren und potentielle Abwanderer rechtzeitig und durch gezielte Angebote weiterhin an das Unternehmen zu binden. Gleichzeitig soll durch eine detaillierte Analyse der Kommunikationswege bestehender Kunden und deren Rolle im sozialen Netz ein Ansatz geschaffen werden, um Werbung zielgerichteter und kundenorientierter zu positionieren. Ziel ist die Anwerbung latenter Neukunden aus dem sozialen Netz bestehender Kunden und damit langfristige Erhöhung des Market Shares des Unternehmens.
Lösung	Als Antwort hat HP eine Social-Network-Analysis-Lösung auf Basis des vom Unternehmen genutzten SAS Customer Link Analysis Tools entwickelt. Der Dateninput wird aus Anruf- und SMS-Verbindungen generiert, die netzintern aber auch zu externen Netzen entstehen. Ausgehend von der Art und Anzahl der Verbindungen der einzelnen Telefonteilnehmer untereinander, identifiziert und clustert die SNA Lösung besonders eng vernetzte Benutzer in sog. »Communities« und analysiert kennzeichnende Eigenschaften. Parallel wird die Beziehung der Kunden zu Nutzern externer Anbieter (OO – Other Operators) und vor kurzem weggewechselten Nutzern (MNP – Mobile Number Portability) analysiert. Im Ergebnis wird ein Risk Score abgeleitet, der auf Basis der Beziehungen zu OO und MNP und unter Beachtung der Community Charakteristika das Risiko eines Anbieterwechsels für jeden Kunden widerspiegelt. Des Weiteren können, indem die Art und Wichtigkeit einer Verbindung zweier Teilnehmer charakterisiert werden, verschiedene Rollen in den Communities identifiziert werden. Diese spiegeln den Einfluss wieder, den Individuen auf die Community und/oder andere Teilnehmer hinsichtlich Anbieterwahl und -wechsel haben. Diese Daten dienen dazu latente, potentielle Kunden zu identifizieren und gleichzeitig durch die Zuordnung zu einer Community zu charakterisieren.
Big-Data-Merkmale	Die hohe Anzahl der Verbindungen, die täglich neu analysiert und zugeordnet werden, sowie die Notwendigkeit, die Daten in nahezu Echtzeit zur Verfügung zu stellen, erfordern eine stark auf Massendaten optimierte Verarbeitung. Dabei. werden derzeit bei über 30 Millionen Subscribern ständig über 500 Millionen Verbindungen zwischen den Teilnehmern ausgewertet.
Nutzen	Bestehende Segmentierungsmodelle können um rollenbasierte Modelle erweitert werden, indem der Einfluss auf das soziale Umfeld durch Leader, Follower etc. verdeutlicht wird. Leader gelten als Kommunikations-Hubs und haben einen starken Entscheidungseinfluss auf ihr Umfeld. Marketingstrategien und Ansätze zur Kundenakquise können durch SNA optimiert werden. Eigenschaften der Communities, Wechsel zwischen den Communities und die Identifikation von Teilnehmern in Schnittstellenbereichen ermöglichen Rückschlüsse auf neue Kundensegmente und Zielgruppen. Der Risk Score zum Abwanderungsrisiko identifiziert Target Customer für Marketingansätze und Werbeangebote, um einem Anbieterwechsel vorzubeugen.

TELECOM ITALIA

10.1.12 (Nº12) Webtrekk GmbH – Realtime-Webanalyse

Anwender

Die Webtrekk GmbH ist ein Anbieter für Webanalyse und fokussiert auf die Themen Online-Marketing und Konversionsratenverbesserung. Der Berliner Spezialist unterstützt Unternehmen, ihre Webseiten und Online-Shops zu optimieren.

Anbieter

Exasol AG, Neumeyerstraße 48, 90411 Nürnberg'
Carsten Weidmann, Head of Presales
Tel: 0911 23 991 0, Email: Carsten.Weidmann@exsasol.com

Problem

Webtrekk verarbeitet bei der Analyse von Webseiten rund 50 Milliarden Datensätze pro Jahr bzw. ca. 40.000 Abfragen pro Tag. Die Kunden erwarten Auswertungen ihres Webcontrollings in Echtzeit, um zeitnahe Entscheidungen treffen zu können.

Das bisher eingesetzte System auf Basis einer MYSQL-Datenbank konnte diese Anforderungen nicht mehr erfüllen. Die Daten mussten voraggregiert werden, wodurch viele Informationen bei den Analysen der Kunden nicht mehr vorhanden waren. Darüber hinaus war auch die Antwortzeit nicht mehr ausreichend: Es konnten keine komplexen Korrelationen in Analysen durchgeführt werden und die grafische Ausgabe der Analysen nahm zu viel Zeit in Anspruch.

Ziel war es daher, eine Lösung zu finden, die sehr komplexe Datenmengen in Echtzeit analysieren, segmentieren und grafisch darstellen kann.

Lösung

Innerhalb von acht Wochen integrierte Webtrekk mit Hilfe des EXASOL-Teams die EXASolution-Datenbank in die bestehende Infrastruktur zur Datenauswertung. Seit Herbst 2008 ist EXASolution bei Webtrekk produktiv im Einsatz. Sämtliche Datensätze der einzelnen Kunden-Webseiten, sei es ein Webshop oder eine herkömmliche Webseite, gelangen in einem ersten Schritt auf sogenannte Trackserver. Im zweiten Schritt werden die Daten von dort zyklisch auf mehrere EXASolution-Datenbanken übertragen. Ein eigener Reportserver holt die Daten dort ab und visualisiert diese. Über Webtrekks Q3-Lösung können die Nutzer letztendlich jederzeit ihre eigenen Berichte oder Ad-hoc-Analysen abrufen.

Big-Data-Merkmale

Gewaltige Datenberge von 50 Milliarden Datensätzen bzw. 40.000 Abfragen am Tag mussten vor der Umstellung aggregiert werden, um sie überhaupt handelbar zu machen. Die Daten bei der Analyse von Webseiten können unterschiedlicher kaum sein: Blog-Einträge, Beiträge aus Portalen oder aus sozialen Medien. Diese komplexen Datenmengen gilt es ad-hoc zu analysieren, segmentieren und grafisch darzustellen, denn die Kunden von Webtrekk erwarten die Auswertungen ihres Webcontrollings in Echtzeit.

Nutzen

Durch die EXASolution Lösung kann das Unternehmen mit seinem Tool Q3 jetzt Rohdaten verarbeiten sowie Teilmengen dieser Daten in Echtzeit berechnen und grafisch darstellen. Dies verschafft dem Berliner Spezialisten ein Alleinstellungsmerkmal im Markt für Webanalyse-Werkzeuge. Insbesondere können auch multivariate Tests zur Optimierung von Webseiten durchgeführt werden.

Für die Kunden von Webtrekk ergibt sich folgender Nutzen: Durch den Einsatz von Q3 lassen sich Datenanalysen für individualisierte Bannerwerbung beschleunigen, Warenkorbanalysen und Berichte schnell abrufen oder es können bestimmte Angebote minutengenau gesteuert werden, so dass Zielgruppen präzise angesprochen werden können.

Lessons learnt

Rohdaten enthalten viele wertvolle Informationen für Analysen, die durch Vorverdichtung der Daten verloren gehen und im weiteren Prozess nicht mehr zur Verfügung stehen. Gleichzeitig sind kurze Antwortzeiten bei Analysen wichtig, um Entscheidungen schnell und rechtzeitig treffen zu können.

Beide Anforderungen müssen sich nicht widersprechen, sondern können durch den Einsatz geeigneter Werkzeuge erfolgreich umgesetzt werden.

10.2 Einsatzbeispiele aus Forschung und Entwicklung

10.2.1 (№13) Mittelständische Unternehmensberatung – Competitive Intelligence: Trend-Analyse im Internet

Anwender	Mittelständische Unternehmensberatung
Anbieter	Empolis Information Management GmbH Martina Tomaschowski (martina.tomaschowski@empolis.com)
Problem	Strategische Entscheidungen zur Unternehmensentwicklung, der Produktinnovation und Marketing erfordern fundierte Informationen als Basis. Das Internet als Informationsquelle bietet hierzu eine überwältigende Menge von Inhalten. Wissenschaftliche Publikationen, Nachrichten, Unternehmensdaten, Produktbeschreibungen und bewertungen, fast alles steht zur Verfügung. Das meiste davon liegt in unstrukturierter, d.h. textueller Form vor. Suchmaschinen helfen bei der Identifizierung potenziell interessanter Inhalte, die dann manuell ausgewertet werden müssen. Mit anderen Worten, all das müsste gelesen werden. Letztlich scheitert dieses Unterfangen, da bei der Menge der Information hierfür keine ausreichenden Ressourcen zur Verfügung stehen. In der Konsequenz werden Entscheidungen unter großer Unsicherheit getroffen, da die entscheidende Information fehlt oder die ausgewählte Information nicht repräsentativ ist.
Lösung	Die Lösung besteht in einem automatisierten Analyseprozess, der sich linguistischer Text-Mining-Verfahren bedient. Ausgehend von einer Domänenontologie wird ein Informationspool aus möglichst vielen, potenziell relevanten Internetquellen gebildet und automatisch analysiert. Das Ergebnis sind Term x Dokument Matrizen, die zur Kolokations- und Korrelationsanalyse genutzt werden. Damit kann ein umfangreicher Textkorpus empirisch erfasst, ausgewertet und auch in seiner Entwicklung über die Zeit beobachtet werden. Die Maschine übernimmt das Lesen und führt dabei Buch. Mit Hilfe verschiedener Visualisierungstechniken, u. a. Matrixbrowser und Netzstrukturen, kann sich der Nutzer die Gesamtheit der Texte erschließen, anstatt sich auf einzelne Dokumente zu verlassen. Dabei werden auch schwache Signale und Trends sichtbar, die bisher verborgen blieben oder nur zufällig entdeckt wurden.
Big-Data-Merkmale	Je nach Domäne der Analyse müssen riesige Textmengen verarbeitet werden, die in den unterschiedlichsten Formaten, Formen und Sprachen vorliegen. Die resultierenden Matrizen haben viele Millionen Dimensionen. Dabei ist die größte Herausforderung die Dynamik, d.h. der Korpus erweitert sich kontinuierlich. Diese Veränderungen müssen beobachtet werden, um neue Trends und Signale möglichst früh aufdecken zu können. Ein konkretes Anwendungsbeispiel unter vielen ist die Analyse der aktuellen Forschungsaktivitäten im Automobilleichtbau. Die dazu relevante Information erstreckt sich über verschiedene Quellen wie Patente, Geschäftsberichte, Fachzeitschriften, Marktanalysen, aber auch zunehmend Social-Media-Quellen (Variety und Volume). Die Linguistik sorgt für eine »sichere« Analyse der Inhalte – dass zum Beispiel auch über komplexe Satzkonstrukte hinweg der Bezug zwischen Technologie und entwickelnder Organisation erkannt wird. Diese Zusammenhänge werden in eine effizient auswertbare Datenstruktur überführt. Die Velocity ergibt sich einerseits aus der Erweiterung des zugrundeliegenden Wissensmodells und der Einbeziehung neuer Quellen, andererseits aus der Änderungshäufigkeit der einbezogenen Quellen, insbesondere im Social-Media-Bereich.
Nutzen	Eine Analyse ohne Unterstützung nimmt mehrere Tage oder gar Wochen in Anspruch. Relevante Quellen müssen nicht nur gefunden und bewertet, sondern auch bei Eignung detailliert analysiert werden. Dabei besteht die Gefahr, relevante Quellen zu übersehen. Das entwickelte System senkt die Bearbeitungszeiten für die Recherche auf 1-2 Tage und erlaubt zudem die periodische Wiederholung der Recherche, da die Detailanalyse entfällt. Die so erhaltene Übersicht ist schneller und kostengünstiger zu erstellen – unter fundierter Einbeziehung relevanter Quellen.

Lessons learnt	Die Ergebnisqualität ist umso höher, je besser die anfängliche Modellbildung ausfällt und umso sorgfältiger die Quellen selektiert wurden. Die kontinuierliche und iterative Anwendung des Verfahrens sollte dabei als kontinuierlicher Verbesserungsprozess gelebt werden.

10.2.2 (№14) Königliches Technologie Institut Stockholm – Realzeit-Analyse für Smarter Cities

Anwender	Königliches Technologie Institut, Stockholm
Anbieter	IBM Research, IBM Schweden
Problem	Verbesserung des Verkehrsmanagements in Stockholm ■ Verkehrs- und Wetterdaten integrieren und analysieren (GPS, Sensoren, Unfall- und Staumeldungen, Videos, …) ■ Probleme prognostizieren und vorbeugen
Lösung	■ IBM InfoSphere Streams ■ Realzeit-Analyse von 100.000en GPS-Daten/Sekunde ■ Ausweichempfehlungen, alternative Routen
Big-Data-Merkmale	Pro Sekunde werden mehr als 250.000 GPS-Daten sowie zahlreiche weitere Daten aus sonstigen Sensoren- und Videosystemen analysiert. Der Wert der Streams-Computing-Lösung steckt in der Anwendung von Realzeit-Analyse und Flexibilität bei der Integration mit dem Sensor-Netzwerk und den GPS-Daten und der effizienten Verarbeitung dieser Daten mit großer Varianz. Nutzen
Nutzen	■ Einbindung in das »Intelligente Verkehrssystem« der Stadt Stockholm ■ 20 % weniger Verkehr ■ 50 % kürzere Fahrzeiten ■ 20 % weniger Emissionen ■ Übertragbarkeit auf andere Kommunen
Lessons learnt	Es konnten neue Einsichten in die Mechanismen von komplexen Verkehrssystemen gewonnen werden.

10.2.3 (№15) University of Ontario – Verarbeitung von Sensordaten medizinischer Überwachungsgeräte

Anwender	University of Ontario Kanada
Anbieter	IBM Canada
Problem	Kanadische Neonatologen haben festgestellt, dass es Fälle von Infektionen bei Frühgeborenen gibt, die durch medizinische Frühindikation ca. 24 Stunden vor dem Ausbruch der gefährlichen Infektion verhindert werden können. Durch die permanente Überwachung bestimmter Symptome und lebenswichtiger Parameter bei Frühgeborenen mit Hilfe modernster, medizinischer Sensorik sollte die Frühgeborenen-Sterblichkeit vermindert werden. Um die kritischen Warnsignale und Parameter-Änderungen des medizinischen Zustandes frühzeitig zu erkennen, müssen permanent tausende von Sensordaten pro Sekunde analysiert werden.
Lösung	Relevante Muster aus den Sensordatenströmen wurde mit IBM Machine Learning Algorithmen und Real-Time Analyse der Datenströme unter Einsatz von InfoSphere Streams identifiziert. Dadurch erzielte man eine Vorwarnung 24 Stunden vor dem kritischen Ereignis und konnte präventive Medikation einsetzen, um lebensbedrohliche Situationen bei den Frühgeborenen abzuwenden.
Big-Data-Merkmale	Die Lösung ist als Big Data zu klassifizieren, weil sehr große, permanente Datenströme der medizinischen Sensoren in Echtzeit auszuwerten sind. Dazu werden Machine Learning Algorithmen im Streaming Mode eingesetzt. Die Streams-Computing-Lösung geht außerdem mit einer großen semistrukturierten Datenvielfalt um und unterstützt damit eine große Varianz in den Datenströmen. Der Wert der Lösung steckt in der Anwendung von Real-Time Analytics, in der flexiblen Integration mit dem Sensor-Netzwerk und der effizienten Verarbeitung der Daten mit großen Volumina und großer Varianz.
Nutzen	Die Lösung unterstützt frühzeitige und gezielte Medikation zur Abwehr lebensbedrohlicher Infektionen und reduziert so die Frühgeborenen-Sterblichkeit.

UOIT
CHALLENGE INNOVATE CONNECT

10.3 Einsatzbeispiele aus der Produktion

10.3.1 (№16) Energietechnik – Überwachung und Diagnose bei komplexen technischen Systemen

Anwender	Anwender aus dem Bereich der Energietechnik (Maschinenbau)
Anbieter	Empolis Information Management GmbH Martina Tomaschowski martina.tomaschowski@empolis.com
Problem	In vielen Bereichen des Maschinenbaus – wie auch in der Energietechnik – wandelt sich zurzeit das Geschäftsmodell von einer reinen Lieferantenrolle hin zum Dienstleister, d.h. Betreiber der Anlage. Dadurch wächst die Bedeutung von Service und Support. Es findet eine Transformation des Service vom Cost-Center hin zum Profit-Center statt. Telematik und Remote Service sind die technologischen Bausteine, die dies ermöglichen und doch auch eine große Herausforderung darstellen. Moderne Produktionsmaschinen sind mit einer Vielzahl von Sensoren ausgestattet. Die überwachten Geräte liefern kontinuierlich Daten über den Zustand der Maschine und ihrer Komponenten. Dazu gehören normale Vorgänge wie das Anfahren oder Stoppen ebenso wie Ereignismeldungen, Warnungen und Fehlercodes.
Lösung	Die gelieferten Daten werden auf relevante Ereignisse hin untersucht, um so den allgemeinen Zustand der Maschine zu ermitteln. Bei einem Vergleich mit der Historie von Instandsetzungsvorgängen – als Folge von Problemen an der Maschine – werden Trends erkannt und ggf. präventive Maßnahmen eingeleitet. Bei Störfällen helfen ähnlich gelagerte Fälle aus der Vergangenheit bei der schnellen Fehlerbehebung.
Big-Data-Merkmale	In diesen Anwendungsfällen liegt ein enormes **Volume:** Jedes Gerät liefert täglich mehrere Gigabyte an strukturierten und unstrukturierten Daten, die vom Remote Service bearbeitet werden müssen. Der Anwender verfügt aktuell über >1000 Installationen entsprechender Geräte. **Variety:** Die Daten unterscheiden sich je nach Gerätetyp im Format und Inhalten. **Velocity:** Die Daten müssen annähernd in Echtzeit nach relevanten Vorfällen durchsucht und verdichtet werden. Bei 1000 Geräten sind dies 3,5 GB pro Minute.
Nutzen	Der bereits erwähnte Wandel des Geschäftsmodells führt dazu, dass die Anzahl der zu überwachenden Systeme rapide wächst. Im Service Center eines Kunden hat sich diese Zahl beispielsweise in diesem Jahr verdoppelt. Eine entsprechende Personalaufstockung ist weder wirtschaftlich und praktisch möglich, da hier hochqualifiziertes erfahrenes Personal benötigt wird. Durch die notwendige Automatisierung der Analyseprozesse werden Fehler schneller behoben, Wartungsvorgänge optimaler geplant und so die Servicequalität bei geringeren Kosten erhöht. Reine Alarmsysteme haben dabei den Nachteil, dass sie zu viele Signale an die Supportmitarbeiter weiterleiten. Üblicherweise sind 80% der Signale Fehlalarme. Erst der Einsatz wissensbasierter Diagnosetechnologie ermöglicht es, die relevanten 20% zu identifizieren und so vorab zu qualifizieren, dass der Support schnell und angemessen reagieren kann.
Lessons learnt	Die Erfahrung zeigt, dass der größte Teil der routinemäßigen Analyse- und Servicefälle automatisierbar ist. Dadurch können sich die Servicemitarbeiter wieder auf die schwierigen Fälle konzentrieren.

10.3.2 (Nº17) Semikron GmbH – Geschäftsprozesse optimieren durch ganzheitliches Mess- und Prozessdatenmanagement

SEMIKRON innovation+service

Anwender	Semikron ist Hersteller von Leistungshalbleiterkomponenten. Das Unternehmen fertigt an 10 Produktionsstandorten sowohl standardisierte Leistungshalbleiter als auch maßgeschneiderte Systeme und Lösungen wie beispielsweise Transistoren, Ansteuerungen, spezielle Kühlungen für Anlagen, Kondensatoren oder auch Controller-Software.
Anbieter	Exasol AG, Neumeyerstraße 48, 90411 Nürnberg Carsten Weidmann, Head of Presales, Tel: 0911 23 991 0 Carsten.Weidmann@exsasol.com
Problem	Während des Produktionsprozesses bei Halbleiterprodukten fallen eine enorme Anzahl von Mess- und Prozessdaten, Materialbewegungen oder Lieferinformationen an. Semikron stand vor der Herausforderung, für sein Spezialgebiet Messdaten-Archivierung, eine flexible Systemlösung zu etablieren, die eine leistungsstarke Speicherung, Auswertung und Rückverfolgung ihrer Messdaten sicherstellt. Das Projekt strebte drei große Ziele an: Archivierung aller qualitätsrelevanten Kennzahlen zu den verkauften Produkten während der vorgeschriebenen Aufbewahrungsfristen, Online-Verfügbarkeit aller archivierten Daten auf Basis von Auftrags- und Artikelnummer mit Recherche nach einzelnen Merkmalen sowie Bereitstellung der Materialbewegungs-und Lieferinformationen zur lückenlosen Rückverfolgbarkeit der Produktionskette.
Lösung	Im Rahmen einer Teststellung (Proof of Concept) wurden zunächst Testdaten von je einer Anlage jedes Fertigungsbereiches in Deutschland in ein vollständig modelliertes Datenmodell eingespielt. Erfolgreiche Testergebnisse sorgten schnell für die gewünschte Optimierung der Semikron-Prozess-Kette, so dass nach und nach weitere Anlagen und Fertigungsbereiche an das Livesystem angebunden wurden. Bis heute sind fünf internationale Standorte und 60 Messanlagen an die Hochleistungsdatenbank EXASolution angeschlossen. Dort stehen die Daten jederzeit für analytische Abfragen und Reports bereit.
Big-Data-Merkmale	Bei Semikron fallen aus den unterschiedlichsten Datenquellen Informationen an. Mess- und Prozessdaten, Daten über Materialbewegungen oder Lieferinformationen. Insgesamt müssen in kürzester Zeit ca. fünf Milliarden Rohdatensätze verarbeitet und ausgewertet werden. Nur so kann der Leistungselektronikhersteller zeitnah auf Marktänderungen und Auffälligkeiten im Produktionsprozess reagieren. Mit der Anbindung weiterer Produktionsstandorte und –anlagen wird das Prozesswissen in der Datenbank zu erhöht, was wiederum eine wachsende Anwenderzahl mit sich bringt.
Nutzen	Mit einem ganzheitlichen Mess-und Prozessdatenarchiv auf Basis von EXASolution kann Semikron seine Produktions- und Qualitätskennzahlen der verkauften Leistungshalbleiter-komponenten innerhalb der vorgeschriebenen Aufbewahrungsfristen zurückverfolgen. Präventive Analysen (Statistical Process Control) sind auf den gewonnenen Messdaten ad-hoc und ohne Informationsverlust möglich. Die Fachbereiche, beispielsweise Prozess- und Qualitätsingenieure in der Fertigung, sind unabhängig von anderen Abteilungen und können ihre Standard- oder Ad hoc-Reports auf Knopfdruck erstellen. Darüber hinaus ist es jetzt möglich, Produktionsnachweise mit allen verfügbaren Daten zu einem einzelnen Bauteil abzurufen.

Lessons learnt	Big-Data-Projekte sind komplex. Oft sind Unternehmen nicht in der Lage, ihre tatsächlichen Datenbestände für die geplanten Projektvorhaben hinsichtlich ihrer Volumenentwicklung abschätzen zu können. Bei Semikron hat sich beispielsweise gezeigt, dass sie von einem viel größeren Datenvolumen ausgegangen sind, als es tatsächlich der Fall war. Bei dem durchgeführten Proof of Concept stellte sich heraus, dass zwar die Vielzahl an Daten, die in den typischen Produktionsprozessen anfallen, sehr hoch ist, nicht aber das Datenvolumen. Eine Teststellung ist somit ein wertvolles Instrument, um letzte Fragen zum Lösungsszenario auf Seiten beider Projektpartner zu beantworten, um letztendlich wertvolle Zeit zu sparen und keine unnötigen Ressourcen zu binden.

10.3.3 (Nº18) Vaillant – Globale Planung und Steuerung bis auf Produktebene

VAILLANT GROUP

Anwender	Marc Stöver – Head of Group IT Consulting, Vaillant
Anbieter	SAP AG
Ansprechpartner	Boris Michaelis SAP Deutschland AG & Co. KG boris.andreas.michaelis@sap.com
Problem	Über die Jahre haben die Anforderungen der 1200 BI-Anwender bei Vaillant kontinuierlich zugenommen. Das Unternehmen nutzt ein integriertes, globales Planungs- und Controlling-System, das bis auf Einzelprodukt-Kundenebene detailliert ist. Allein der Materialfluss hat in den letzten Jahren rund eine Milliarde Datensätze in den Datenwürfeln produziert. Die Historie jedes Gerätes ist verfügbar. Das Management der verschiedenen IT-Systeme gestaltete sich immer schwieriger und aufwändiger.
Lösung	Vaillant setzt nun SAP-HANA-In-Memory-Technologie und SAP-Analytics-Lösungen produktiv ein. Es wurden verschiedene IT-Systeme zusammengeführt. Somit hat sich auch die Systemlandschaft vereinfacht.
Big-Data-Merkmale	Kombination von Materialien, Kunden, Anzahl von Datensätzen aus dem Rechnungswesen in Verbindung mit Kostenstellen und Profit-Centern usw. sind Basis der Unternehmensplanung und -steuerung[67] und bedürfen daher leistungsfähigster Systeme, um in Echtzeit zu arbeiten. Forecast und Strategieprozesse müssen in kürzester Zeit bearbeitet werden. Die Geschwindigkeitsverbesserungen lagen bei der Datenbereitstellung bei einem Faktor von 4, in den Planungsanwendungen bei 10 und im Reporting bei 60.
Nutzen	■ Höhere inhaltliche Qualität, Aktualität und Homogenität der Daten und Anwendungen in Verbindung mit entsprechender Detailtiefe, von z. B. dem einzelnen Material bis zum EBIT in der Bilanz der Unternehmensgruppe. ■ geringere Fehleranfälligkeit ■ TCO-Einsparungen: ■ zwischen 33% und 66% für kostengünstigere Hardware (keine SSD Speicher und High-end UNIX-Systeme mehr nötig und Ersatz z. B. durch Intel Linux) ■ deutliche Reduzierung der Support- bzw. SLA- und Beratungskosten im Systembetrieb um 21% sowie Optimierung des Datenmanagements inklusive Archivierungs- und Back-up-Konzepte ■ Reduzierung der Datenbereitstellung um Faktor 4 bzw. Faktor 10 bei Planungsanwendungen ■ Grundlagen für den weiteren Ausbau von Predictive-Analytics-Szenarien sind gelegt.
Lessons learnt	Allein die Umstellung der Systemlandschaft auf innovative Big-Data-Architekturen aus technischer IT-Perspektive ergeben belastbare Business Cases zur Reduzierung des TCO. Noch deutlich übertroffen werden für Fachabteilungen die Resultate aus dem Mehrwert der neuen Lösungen und Möglichkeiten in Verbindung mit der drastischen Reduzierung der Bearbeitungszeiten durch die Anwender.

[67] Pro genannter Kategorie sind viele Tausend Datensätze zu verarbeiten, mitunter auch Zahlen im zweistelligen Millionenbereich.

10.4 Einsatzbeispiele aus Service und Support

10.4.1 (№19) Automobilhersteller – Ganzheitliche Qualitätsanalyse durch integrierte Daten	
Anwender	Anonymisiert
Anbieter	T-Systems International GmbH
Problem	Nur wenige Experten haben einen wirklich kompletten Überblick über alle verfügbaren Daten zu einem Fahrzeug – von der Entwicklung über die Produktion bis hin zum After-Sales-Service. Es gibt keine umfassende Verantwortlichkeit für alle Daten entlang der gesamten Wertschöpfungskette. Heterogene IT-Umgebungen sorgen immer wieder für zeitraubende Datenrecherchen. Die Bereitstellung und Analyse dieser Informationen kann nachhaltig zur Qualitätssteigerung sowie Fehlerfrüherkennung führen. Allerdings nimmt die anfallende Datenmenge durch immer neue und komplexere elektronische Komponenten zu, so dass eine zeitnahe Bereitstellung der Daten eine Herausforderung darstellt.
Lösung	■ Schaffung einer einheitlichen Schnittstelle für die mehr als 2.000 internen Nutzer bei verschiedensten Funktionen und Analysen ■ Sammlung aller Datenquellen in einem zentralen Data Warehouse ■ Herstellung der Konsistenz aller Daten und klarer Verantwortlichkeiten ■ Optimierung der Analyseunterstützung mit integrierten Frühwarnsystemen ■ Standardisierung der Analyse und des Reportings in den Bereichen After Sales und Technologie, um Synergien zu schaffen.
Big-Data-Merkmale	**Volume:** ca. 10 Terabyte mit deutlich steigender Tendenz. **Velocity:** Kontinuierliches Datenwachstum sowie die permanente Aktualisierung der Daten führen zu einem hohen Bedarf für hoch performante Abfragen. **Variety:** Dauernde Anpassung der Dateninhalte durch neue Software-Stände sowie immer neue elektronische Steuerkomponenten führen zu neuen Datenstrukturen. Die umfangreichen technischen Daten liegen oftmals nur semistrukturiert vor.
Nutzen	■ Den Entscheidungsträgern stehen nun alle Qualitätsinformationen über die Fahrzeuge sowie die korrespondierenden Analysewerkzeugen zur Verfügung. ■ Steigerung der Kundenzufriedenheit ■ Profitabilitätssteigerung durch Fehlerfrüherkennung ■ Qualitätssteigerung durch Ausschließen von Fehlerquellen ■ Homogene Datenquellen, allseits verfügbare Daten sowie die Integration von Analyse, Textmining und Reporting verkürzen die Zeiten für die Fehlererkennung und korrektur bei den Fahrzeugen.
Lessons learnt	■ Variety ist die größte Herausforderung, da sich die Datenstrukturen häufig ändern oder erweitert werden. Transparente Prozesse sind notwendig, um die Datenvielfalt zu beschränken. ■ Data Governance bietet die Grundlage für ein transparentes und strukturiertes Datenmanagement. ■ Die mit Volume und Velocity verbundenen Herausforderungen können durch Hardware- und Softwareoptimierungen bewältigt werden, während dem die durch ständig neue Anforderungen gekennzeichnete Variety durchaus kritisch werden kann.

10.4.2 (Nº20) Treato – Analyse von unstrukturierten Gesundheitsdaten



Anwender	Treato / First Life Ltd. 9 HaAtsma'ut Road Yehud, ISRAEL 56300 info@treato.com
Anbieter	Cloudera Inc., 220 Portage Avenue, Palo Alto, CA 94306, USA cloudera@bhavacom.com
Problem	Im Internet tauschen Patienten und Ärzte in tausenden von Foren und Blogs ihre Erfahrungen mit Behandlungsmethoden von Krankheitsbildern aus. Das Sammeln und Analysieren dieser unstrukturierten Daten war bisher nur mit sehr hohem personellen Einsatz möglich. Zudem fehlte es einzelnen Personen an ausreichendem Fachwissen, um aus den Daten verwertbare Ergebnisse abzuleiten. Technisch waren der Auswertung Grenzen gesetzt, da nicht nur die Masse an Daten kaum zu bewältigen war, sondern auch die verwendete Sprache in den Foren. Da die meisten Patienten keine medizinische Terminologie verwenden, lassen sich die reinen Textinformationen kaum gezielt abgreifen. Erste Versuche machte Treato mit einer Technologie, mit der sich die Verarbeitung von Massendaten nicht durchführen ließ. Daher nutzte Treato nur die Daten einiger Dutzend Webseiten. Doch zeigten sich schon hier Erfolge. Beispielsweise berichtete eine große Zahl von Patienten über die Nebenwirkungen eines Asthma-Medikaments, vor denen offizielle Stellen erst vier Jahre später warnten. Treato suchte daher nach Big-Data-Analysemethoden, mit denen sich tausende Datenquellen schnell und zu geringeren Kosten verarbeiten lassen.
Lösung	Treato greift Posts aus dem Internet ab, analysiert sie mit einer Big-Data-Infrastruktur auf Basis von Hadoop-Technologie von Cloudera und verarbeitet so heute bis zu 200 Millionen Posts pro Tag. Die Lösung erfüllt vier Kriterien: ■ Zuverlässige und skalierbare Speicherung ■ Zuverlässige und skalierbare Auswertungsinfrastruktur ■ Ausreichende Suchmaschinenkapazitäten, um Posts mit hohem Nutzwert finden zu können ■ Skalierbare Echtzeitspeicher, um Statistiken mit hohem Nutzwert erstellen zu können.
Big-Data-Merkmale	**Volume:** Die Lösung verarbeitet pro Tag bis zu 200 Millionen Posts im Internet. Ohne Big-Data-Technologie konnte Treato nur wenige rund zehn Millionen Post verarbeiten. **Variety:** Die ausgewerteten Daten sind in mehrfacher Hinsicht vielfältig. Sie sind unstrukturiert, da es sich um frei formulierte Posts von Millionen von Internet-Nutzern handelt. Sie sind sprachlich stark unterschiedlich, da sie nur wenig Fachterminologie enthalten, aber in Zusammenhang mit fachlicher Terminologie gebracht werden müssen. Und sie bestehen aus einer Mischung aus Text und Zahlen. **Velocity:** Es können pro Tag mehrere hundert Millionen Informationen aus Internet-Seiten ausgelesen und verarbeitet werden. Die Analysen stehen ad hoc zur Verfügung.
Nutzen	Die Analysen helfen zum Beispiel herauszufinden, welche Nebenwirkungen eines Medikaments Patienten am häufigsten beschreiben oder mit welcher Behandlung sie für vergleichbare Erkrankungen zufrieden sind. Bisher dauerte es oft mehrere Jahre, bis Behörden ein Medikament mit starken Nebenwirkungen vom Markt genommen haben oder bis ein erfolgreiches Medikament für bis dahin nicht bekannte Krankheitsbilder zum Einsatz kam.
Lessons learnt	Dieses Big-Data-Einsatzbeispiel zeigt, wie das unstrukturierte, weltweite Crowd-Wissen von Endverbrauchern aus Foren, Blogs und anderen Social-Media-Kanälen strukturiert analysiert und anderen Verbrauchern zur Verfügung gestellt werden kann. Die Analysen lassen sich auch von Herstellern für die Weiterentwicklung von Produkten nutzen.

10.5 Einsatzbeispiel aus Distribution und Logistik

10.5.1 (№21) TomTom Business Solutions – Flottenmanagement in Echtzeit

Anwender	TomTom Business Solutions ist eine spezielle Sparte von TomTom NV für den Bereich kommerzieller Fahrzeugflotten. Das 2005 gegründete Unternehmen gehört mit »TOMTOM WEBFLEET« zu den weltweiten Marktführern und wächst unter den Anbietern von Telematikdiensten in Europa am schnellsten. TomTom-Lösungen werden in mehr als 175.000 Fahrzeugen eingesetzt.
Anbieter	Fujitsu Technology Solutions GmbH Robert.Guzek@ts.fujitsu.com
Problem	Rund 1,5 Milliarden in Echtzeit eingehende Meldungen und mehr als 1 Milliarde Anfragen pro Monat bilden eine wachsende Datenflut und sind zu bewältigen.
Lösung	Für die Kundenanforderung – 200.000 Input-Output-Operationen pro Sekunde (IOPS) und Antwortzeiten unterhalb einer Millisekunde – wurde eine abgestimmte Hard- und Software-Lösung entwickelt. Zum Einsatz kam Oracle mit ASM in einer Linux-basierten Serverfarm, AIS Connect, Cisco Switche, ETERNUS SF Software für Überwachung und Management sowie ETERNUS 8400 Storage Systeme.
Big-Data-Merkmale	Über 175.000 Geschäftsfahrzeuge senden im Minutentakt wahre Datenberge von ihren »Connected Navigation Devices«. Sie übermitteln Statusinformationen, Positionsangaben, Daten von digitalen Tachographen, Verbrauchswerte, Orderdaten sowie weitere Informationen und teilen exakte Ankunftszeiten mit – alles in Echtzeit. Täglich kommen auf diese Weise über 70 Millionen Meldungen zusammen. Sie müssen umgehend verarbeitet werden, denn Disponenten und Fuhrparkmanager bei den Kunden verlangen nach Realtime-Informationen für die Steuerung ihrer Flotten.
Nutzen	Gewährleistung hochverfügbarer Realtime-Prozesse im Big-Data-Umfeld, gemäß den Anforderungen zur Bewältigung der wachsenden Datenflut von rund 1,5 Milliarden in Echtzeit eingehenden Meldungen und von mehr als 1 Milliarde Anfragen pro Monat. Kernsegmente hierfür sind umfassender Datenschutz, robuster und zuverlässiger Betrieb, hohe Ausfallsicherheit durch redundante Auslegung, geringer Energieverbrauch und einheitliche Speicherverwaltung sowie flexible Ausbaumöglichkeiten bei wachsenden Leistungsanforderungen.
Lessons learnt	Um die kompletten Anforderungen des Kunden in Big-Data-Projekten bedienen zu können. ist übergreifendes Know-how erforderlich, das die Konfiguration von Hard- und Software, das Tuning und technisches Consulting umfasst.

10.6 Einsatzbeispiele aus Finanz- und Risiko-Controlling

10.6.1 (Nº22) Europäischer Spezialist für Kreditkartensicherheit – Kreditkartenbetrugsanalyse

Anwender	Europäischer Spezialist für Kreditkartensicherheit
Anbieter	Fraunhofer Institut für Graphische Datenverarbeitung (IGD), Darmstadt, Deutschland Jörn Kohlhammer, Abteilungsleiter »Information Visualisierung und Visual Analytics« joern.kohlhammer@igd.fraunhofer.de
Problem	Kreditkarten sind auch in Deutschland ein immer häufiger genutztes Zahlungsmittel. Wesentlich für das Vertrauen in elektronische Zahlungsmittel ist ein effektiver Schutz vor Missbrauch. Dafür müssen Banken und Kreditinstitute neue Betrugsstrategien möglichst früh erkennen, um Gegenmaßnahmen zu entwickeln. Aktuelle Reklamationen der Kunden werden daraufhin untersucht, ob dahinter irgendein Muster steckt, das auf eine Betrugsstrategie hindeuten könnte.
Lösung	Hinter dieser Aufgabe stehen mindestens drei große Herausforderungen: ■ Die erste ist die schiere Menge der täglich anfallenden Transaktionen[68]. ■ Die zweite Herausforderung besteht darin, dass die Betrüger einen hohen Aufwand treiben, ihre Spuren in den Daten zu verwischen. Nach welchen Zusammenhängen gesucht werden muss, ist daher nicht immer sofort klar. Ob es der Zeitpunkt der Transaktion ist, die Region, oder ob auch der genaue Ort eine Rolle spielt, kann erst im Vergleich mit anderen Betrugsfällen und mit korrekten Transaktionen nachgewiesen werden. ■ Die dritte Herausforderung ist, dass das Wissen um eine Betrugsstrategie schnell veraltet. Sobald die Gegenmaßnahmen greifen, werden neue Strategien entwickelt. Aktuelle Entwicklungen müssen die Betrugsanalysten der Bank daher ständig im Blickfeld behalten. Durch die Kombination von automatischer und interaktiver Datenanalyse können gerade sich dynamisch verändernde Muster effektiv untersucht werden. Betrugsstrategien, für die es noch keine Regel gibt, werden allein mit automatischen Verfahren möglicherweise nicht gefunden. Der Mensch kann jedoch bei der Sichtung der Daten ungewöhnliches Verhalten identifizieren, das anschließend automatisch in Alarmierungsregeln umgewandelt wird. Das Fraunhofer IGD entwickelt dafür analytische Visualisierungstechniken, mit denen auch komplexe Muster leicht gefunden und untersucht werden können.
Big-Data-Merkmale	**Volume:** Jede der Millionen Transaktionen pro Tag enthält 25 und mehr Attribute (Höhe der Transaktionen, Ort, Währung etc.), die relevante Hinweise auf einen Betrug liefern können. **Variety:** Dabei sind diese Attribute aber selten isoliert – sondern erst in Kombination – interessant[69]. Das Ergebnis ist eine hohe Komplexität von immer wieder anders ausgeprägten Betrugsmustern. **Velocity:** Ein weiterer Big-Data-Aspekt ist der hohe Zeitdruck, um nach der ersten Reklamation effektive Alarmierungsregeln zur Abstellung des Betrugsmusters zu finden.
Nutzen	Die Analyse sich dynamisch verändernder Muster kann nicht allein durch automatische Verfahren gefunden werden. Hier kann der Mensch mit seinen eigenen Stärken – Erfahrungswissen und Mustererkennung – eingreifen, um im Ernstfall präzisere Zusammenhänge schneller zu erkennen.
Lessons learnt	Ähnliche Ansätze sind auch relevant für die Untersuchung technologischer oder wirtschaftlicher Trends etwa aus Internetquellen. Auch diese sollen möglichst früh erkannt werden, aber gerade im frühen, noch interessanten Stadium ist die Faktenbasis oft noch zu klein, um Ergebnisse zu finden, die signifikant genug für automatische Verfahren sind. Techniken, die die Stärken des Menschen für die Analyse nutzen, müssen dabei aber möglichst so gestaltet werden, dass sie sich nahtlos in dessen Arbeitsabläufe und Strukturen einpassen.

[68] 2009 wurden 7,2 Mrd. VISA-Kreditkarten-Transaktionen registriert.

[69] z.B. eine bestimmte Transaktion zu einer bestimmten Tageszeit

10.6.2 (Nº23) Paymint AG – Fraud Detection in Kreditkartentransaktionen

Anwender	Paymint AG für mehrere deutsche und internationale Banken und Kreditkartenprozessoren.
Anbieter	Fraunhofer IAIS in Zusammenarbeit mit Paymint AG Eschborner Landstraße 55 60489 Frankfurt am Main
Problem	Kreditkartenbetrug hat sich zu einem großen Problem im kartengestützten Bezahlen entwickelt. Die weltweiten Kosten von Kreditkartenbetrug sind im letzten Jahr auf mehr als 10 Milliarden Euro weltweit gestiegen. Da Betrüger immer schneller reagieren und existierende Betrugserkennungs-Tools überwinden, ist damit zu rechnen, dass diese Kosten auch weiterhin steigen werden. Der Schlüssel zu einem effektiven Fraud Management liegt in der schnellen Reaktion auf neu aufkommende Betrugsmuster.
Lösung	Die PAYMINT AG hat sich in den vergangenen beiden Jahren als Spezialist für die Vermeidung von Kartenmissbrauch etabliert. In Zusammenarbeit mit Fraunhofer IAIS wurde eine Software zur automatischen Regelerstellung auf Basis neuester und effizienter Fraud-Mining-Technologie erweitert: MINTify rule. Neue Fraud-Szenarien und -Muster werden selbständig erkannt und auf deren Basis werden automatisch transparente und nachvollziehbare Regeln zur Verhinderung des neuen Betrugsmusters gebildet.
Big-Data-Merkmale	In der Analyse von Kreditkartentransaktionen muss ein hoher Volumen von Transaktionen bearbeitet werden, in großen, Unternehmen sind mehrere Milliarden Transaktionen pro Monat üblich. Innerhalb dieser Menge muss die kleine Teilmenge der betrügerischen Transaktionen – meist weniger als 0,1% - erkannt und treffsicher beschrieben werden. Einzelne Muster können sich auf eine komplexe Historie vergangener Transaktionen mit mehreren Tausend Merkmalen beziehen, d.h. es muss eine große Vielfalt an potentiell interessanten Mustern durchsucht werden. Zusätzlich kann sich die Datenlage jederzeit kurzfristig ändern, wenn Betrüger neue Strategien versuchen und wiederum Gegenmaßnahmen eingeleitet werden.
Nutzen	MINTify rule wurde kurz nach der Fertigstellung bereits erfolgreich bei einem führenden europäischen Zahlungsabwickler implementiert und schützt dort ein Portfolio von vielen Millionen Kreditkarten. Durch den Einsatz von Big-Data-Technologien konnte die Zeit für die Erstellung eines neuen Regelsatzes zur Betrugserkennung sowie der manuelle Aufwand bei der Bearbeitung neuer Fälle klar reduziert werden. Dies ermöglicht eine schnellere Reaktion auf sich neu entwickelnde Betrugstrends - mit den offensichtlichen finanziellen Auswirkungen.
Lessons learnt	Die drei Merkmale von Big Data – Volume, Variety, und Velocity – sind eine große Herausforderung. Gerade im Data Mining wird dies deutlich, wenn komplexe Muster in großen Datenmengen in Realzeit gelernt werden sollen.

10.6.3 (Nº24) United Overseas Bank (Singapur) – Risikoabschätzung in Echtzeit

Anwender	United Overseas Bank (Singapur) (erste europäische Banken in Pilotprojekten)
Anbieter	SAS Institute
Problem	Als Entwicklungspartner von SAS hat die United Overseas Bank ein High-Performance Risk-Produkt von SAS getestet. Die Berechnung aller vorhandenen Risiken im Portfolio einer Bank ist fachlich wie technisch anspruchsvoll. Die Risiken sind in diesem Fall verteilt auf etwa 45.000 verschiedene Finanzinstrumente und werden bestimmt über etwa 100.000 Marktparameter (Preise, Fristen, Fälligkeiten, etc.). Die Berechnung des Gesamtrisikos setzt etwa 8,8 Mrd. einzelne hochkomplexe Value-at-Risk-Berechnungen voraus . Im Rahmen eines Stresstests sollen nun Marktauswirkungen auf das Gesamtrisiko der Bank untersucht werden. Mittels umfangreichen fachlichen Know-hows kann die Aufgabe zwar gelöst werden, dauert aber etwa 18 Stunden. Eine zeitnahe Reaktion auf neu eintretende Marktrisiken ist damit nicht möglich.
Lösung	Durch den Einsatz von In-Memory-Technologie und speziell auf höchste Parallelität der Analytik optimierte Risikoberechnung einer neuen SAS High-Performance Analytics-Lösung konnte die Dauer zur Berechnung auf wenige Minuten gesenkt werden. In aktuellen Tests ist es sogar möglich, jeweils inkrementell die Auswirkung von neuen Marktinformationen auf das Gesamtrisikoportfolio zu berechnen.
Big-Data-Merkmale	Das Volumen der einzubeziehenden Daten ist nicht so entscheidend, wie der Aufwand, der für die Berechnung getrieben werden muss. Das hat bisherige Rechensysteme überfordert. Durch den Einsatz von In-Memory-Technologie und Complex Event Processing ist es nun möglich, schnell sich ändernde Parameter quasi in Echtzeit mit in die aufwändige Analytics einzubeziehen.
Nutzen	Die bisherige Pflichtübung, zum Beispiel auf Drängen der Aufsichtsbehörden, kann nun im operativen Geschäft genutzt werden, indem Handelsstrategien im Voraus geprüft und neue Marktereignisse in ihren Wirkungen schneller eingeschätzt werden können.
Lessons learnt	Entscheidend ist das Denken in Geschäftsprozessen. Wird nur ein Teil beschleunigt – ganz extrem in diesem Fall –, der Gesamtprozess bleibt aber unangetastet, so lässt sich der Vorteil nicht realisieren. Sowohl das Datenmanagement im Vorfeld, als auch die Echtzeit-Nutzung der Echtzeit-Ergebnisse sind bestimmende Faktoren für den erfolgreichen Einsatz dieser neuen Lösung.

UOB

10.7 Einsatzbeispiele aus Administration, Organisation und Operations

10.7.1 (№25) Aadhaar-Projekt – Personenidentifikation indischer Bürger als Grundlage für Verwaltungs- und Geschäftsprozesse

Anwender
UIDAI (Unique Identification Authority of India)
Aadhaar-Program

Anbieter
Projekt der öffentlichen Hand mit Unterstützung von Firmen wie Wipro Technologies mario.palmerhuke@wipro.com, HP India und anderen.[70]

Problem
»The Unique Identification Authority of India (UIDAI) has been created as an attached office under the Planning Commission. Its role is to develop and implement the necessary institutional, technical and legal infrastructure to issue unique identity numbers to Indian residents. [...] Unique identification project was initially conceived by the Planning Commission as an initiative that would provide identification for each resident across the country and would be used primarily as the basis for efficient delivery of welfare services. It would also act as a tool for effective monitoring of various programs and schemes of the Government.«[71]
Erste Diskussionen und Entscheidungen gab es bereits im Jahre 2006. Seitdem wurde kontinuierlich an Definition und Umsetzung gearbeitet. Das Project wird federführend von einer Regierungsbehörde mit Hilfe von privaten und öffentlichen Partnern umgesetzt.
Das Ziel ist eine zentrale Datenbank für jeden indischen Staatsbürger, die sowohl eine Identifikations-ID wie auch andere Daten enthält. Zum Zweck der Identifikation werden auch biometrische Daten gespeichert.

Lösung
Die Lösung wird zum überwiegenden Teil unter Nutzung öffentlich verfügbarer Software wie Hadoop (Big Data), Rabit MW, Mule ESB u.a. implementiert.

Big-Data-Merkmale
Mit Identifikationsdaten für 1.2 Milliarden Einwohner, die zudem auch noch biometrische Daten enthält, stößt das Aadhaar-Projekt in neue Dimensionen für diese Art von Anwendungsfall vor. Zudem sollen die Daten de facto in Echtzeit abfragbar sein, um die Identifizierung für öffentliche, aber auch kommerzielle Geschäftsprozesse nutzen zu können.
Die Ausgabe der UIDs hat mittlerweile begonnen und es ist das kommunizierte Ziel, in den ersten 5 Jahren 500+ Millionen UIDs mit den notwendigen Verifizierungen und Daten zu kreieren.

Nutzen
- Das Ziel von Aadhaar ist die Einführung eines einzigen, zentralen und von der Regierung verifizierten Identifikationsservices. Indischen Staatsbürgern soll das wiederholte Vorlegen von Identifikationsdokumenten z. B. für die Eröffnung eines Bankkontos oder bei der Beantragung öffentlicher Leistungen wie z. B. Fahrerlaubnis oder Reisepass erspart werden.
- Durch den eindeutigen Identifikationsnachweis wird insbesondere den armen und unterprivilegierten Einwohnern überhaupt der Zugriff auf das Bankensystem sowie öffentliche und private Unterstützungsleistungen gewährt, da u.a. das Missbrauchsrisiko stark reduziert wird.
- Bundesstaatübergreifende Mobilität für Migranten durch die Zentralisierung der Identifikation.
- Möglichkeit der Personenidentifikation »anytime, anywhere, anyhow« durch einen zentral kontrollierten, vor Missbrauch geschützten Service der Regierung

Lessons learnt
- »Best Practices« für die Durchsatz- und Startzeitoptimierung von Hadoop-Clustern
- Arbeitsstandards zur Definition und Modellierung von biometrischen und demographischen Daten in Big Data Technologien wie Hadoop.

Information
http://uidai.gov.in/

70 Eine vollständige Liste der vergebenen Aufträge findet sich unter http://uidai.gov.in/contracts-awarded-link.html.
71 Vgl.: http://uidai.gov.in sowie http://uidai.gov.in/why-aadhaar.html

10.7.2 (№26) Anbieter für Dialog-Marketing per E-Mail – Plattform für Versand vertrauenswürdiger E-Mails

Anwender	Führender Anbieter für Dialog-Marketing per E-Mail
Anbieter	fun communications GmbH, Karlsruhe, Deutschland johannes.feulner@fun.de
Problem	Spam- und Phishing-Mails sind für kommerzielle E-Mail-Versender mehr als nur ein Ärgernis – die Unternehmensklientel wird von Dritten unter dem eigenen Namen attackiert. Während Spam-Mails meist anhand ihrer Inhalte aussortiert werden können, entsprechen Phishing-Mails möglichst genau den Original-Mails, um die Spam-Filter zu umgehen. Der Kunde beauftragte fun communications aus diesem Grund mit der Konzeption und der Realisierung eines Dienstes, der gewerbliche Versender von E-Mails, die sich in einer speziellen Vertrauensgemeinschaft organisieren, auch als solche identifiziert und entsprechend ausweist. Die Empfänger sollen bereits in der Betreffzeile einer E-Mail und somit auf den ersten Blick erkennen, dass es sich bei dieser Mail nicht um eine Spam- oder Phishing-Mail, sondern um eine »echte« E-Mail mit seriösem Inhalt handelt. Auf der Plattform muss der sichere Betrieb mit dutzenden E-Mail-Versendern und zahlreichen E-Mail-Providern gewährleistet werden. Das rasant steigende Transaktionsvolumen von inzwischen über 1 Mrd. Ereignissen pro Monat fällt verteilt bei den E-Mail-Providern an und muss permanent zur Qualitätssicherung, Abrechnung und Performanceanalyse ausgewertet werden.
Lösung	Unter Verwendung von DomainKeys Identified (DKIM) Signaturen entwickelte fun communications eine Plattform, die auf unterschiedlichste Art in den Mailprozess eines E-Mail-Providers integriert werden kann, um auf Basis dieser Signaturen Aussagen über die Reputation der Versender zu treffen. Dem E-Mail-Empfänger wird die Echtheit der Mails im eigenen Client durch die Anzeige eines entsprechenden Prüfsiegels und des Logos des Versenders signalisiert – er sieht sofort, dass eine Mail geprüft wurde und damit authentisch ist. Zentrale Komponente der Plattform ist ein Hadoop-Cluster zur Datenhaltung und Auswertung der Event-Logs der E-Mail-Provider. Zudem müssen zu jedem beliebigen Zeitpunkt neue Versender sowie neue Provider im laufenden Betrieb integriert werden können. ■ 7 Hadoop-Knoten auf Standard-Hardware verarbeiten das Monatsaufkommen von 1 Mrd. Events in weniger als 3 Stunden ■ Zentrale Konsolidierung von Events verschiedener E-Mail-Provider ■ Unsortierter und unregelmäßiger Dateneingang ■ Stete, zeitnahe Datenauswertung ■ Integration neuer Mail-Versender und neuer Provider im laufenden Betrieb ■ Hadoop mit MapReduce und Hive hat die Performance gegenüber einer Lösung mit relationaler Datenbank um den Faktor 50 erhöht ■ Zusätzliche Auswertungen, Probeläufe und Wartungszeiten jederzeit möglich ■ Plattform auf zukünftig zu erwartete Last ohne zusätzliche Hardware konzipiert.
Big-Data-Merkmale	**Volume:** Die Anwendung ist im Markt sehr erfolgreich und durch ein steiles Ansteigen der Datenmenge gekennzeichnet. Aus 150 Mio. Events pro Monat sind innerhalb weniger Monate inzwischen monatlich über 1 Mrd. Events geworden. Das Wachstum hält an und beschleunigt sich weiterhin. **Variety:** Die Daten fallen verteilt bei angeschlossenen E Mail-Providern an. Unterschiede im Datenformat, der Datenqualität und auch in der zeitlichen Bereitstellung müssen von der Plattform kompensiert werden. **Velocity:** Eine erste Lösung mit einer klassischen relationalen Datenbank stieß trotz starker Optimierung schnell an ihre Grenzen. Die Verarbeitung und Auswertung der Daten eines Tages dauerte schließlich länger als 24 Stunden. Die Lösung mit dem Hadoop-Cluster brachte die Performance auf ca. 2 Stunden und kann jetzt über zusätzlichen Hardware-Einsatz auch bei wachsendem Datenvolumen konstant gehalten werden.

Nutzen	Der Einsatz von Big-Data-Technologien wie Hadoop und Hive hat gegenüber dem Einsatz einer relationalen Datenbank bei gleicher Hardware eine Steigerung der Performance um den Faktor 50 gebracht. Die weitere Skalierung ist auch bei extremem Wachstum weiterhin mit günstiger Standard-Hardware möglich. Inzwischen wurde die Plattform mit Splunk (www.splunk.de) anstelle von Hadoop nochmals implementiert. Während man bei Hadoop klassische Softwareentwicklung betreibt, wird Splunk, ein Out-of-the-box Werkzeug, durch einfaches Customizing angepasst. Die Performance erwies sich als ähnlich gut wie die Hadoop-Lösung. Splunk bringt aber noch weitere Vorteile: ■ Entwicklungszeit um Faktor 5 reduziert (keine Programmierung) ■ Exploration der Daten mit neuen Fragestellungen ist jederzeit in Echtzeit möglich ■ Grafische Reports auch personalisiert für die Plattformkunden mit minimalem Aufwand erstellt.
Lessons learnt	Big-Data-Aufgaben, an denen herkömmliche Technologien scheitern, lassen sich mit den aktuellen Technologien bewältigen und versprechen gute Skalierbarkeit für weiterhin steigende Anforderungen. Die Auswahl geeigneter Werkzeuge kann die Entwicklungszeiten dramatisch verkürzen.

10.7.3 (№27) Paketdienstleister – Sicherung der Compliance

Anwender	Anonymisiert
Anbieter	T-Systems International GmbH
Problem	Die Sicherung der Compliance bei der Archivierung bildet in vielen Branchen eine wachsende Herausforderung. Das gilt auch für einen führenden globalen Paketdienstleister: Bei der Zustellung von Paketen sind eine revisionssichere Langzeitspeicherung aller Dokumente zu ein- und ausgehenden Sendungen – dazu gehören elektronische Unterschriften auf Handheld bei Paketempfang wie auch gescannte Empfangsbelege von Poststellen – sowie der Zugriff über ein Web-Portal notwendig.
Lösung	Aufbau eines unternehmensweiteren Archivs, das alle ein- und ausgehenden Paketlieferungen dokumentiert ■ Implementierung des Archivs auf Basis von T-Systems Image Master ■ Betrieb im Hochverfügbarkeits-Rechenzentrum der T-Systems ■ Erfüllung aller regulatorischen Vorgaben (Compliance).
Big-Data-Merkmale	**Volume:** Das enorme Volumen sowie das Dokumentenwachstum erfordert eine hoch performante Achivierungslösung: ■ Archivgröße: 19 TB oder 5 Mrd. Dokumente ■ Wachstum/Monat: 550 GB oder 80 Mio. Dokumente (à 10 Jahre Aufbewahrungspflicht) ■ Archivierungslast (Messung über 24h Dauerarchivierung): ■ 17,4 Mio. Dokumente ■ 34,8 Mio. physikalische Dateien ■ 51,1 Mio. Datensätze **Velocity:** Der enorme Zuwachs der Dokumente sowie schnelle Reaktionszeiten bei konkreten Nachfragen erfordern sehr kurze Such- und damit verbunde Antwortzeiten des Archivs. Antwortzeiten (End-2-End-Messung zum Retrieval eines Dokuments): ■ Minimum 0,94 Sek. ■ Durchschnitt 1,73 Sek. **Variety:** Der Kunde kann variabel Dokumentart und Ablage bestimmen (SOA Backbone – Unternehmens-Archivbus). Dabei kann ImageMaster diese frei als eigene Instanz, Mandant, Datenbank-Schema oder Ordner einer bestehenden Ablage/Archiv verwalten. Image Master erkennt die Eingangsströme anhand von Attributen und kann damit komplexe Dokumentenverarbeitung mit Konvertierung, Zerlegung, Attributextraktion und Ressourcenversorgung unterschiedlichster Formate und Dokumentenklassen sowie unterschiedlichste Archivierungsströme gewährleisten.
Nutzen	■ Umfangreiche Compliance-Anfragen können auf Basis der Plattform effektiv bedient werden. ■ Transparenz und ständiger Zugriff auf die Paketzustellungen der letzten 10 Jahre. ■ Guter Überblick über die geleisteten Paketzustellungen, was bei Bedarf für fundierte Analysezwecke benutzt werden kann.
Lessons learnt	■ Bei Enterprise Backbone-Archiv-Projekten ist es von besonderer Bedeutung, sich auf ein robustes und über Zeit und Volumen skalierbares Datenbank- und IT-Konzept stützen zu können. ■ Spätestens alle fünf Jahre muss eine Soft Migration der Archive und Review der Datenablagekonzepte durchgeführt werden, da sich Unternehmen immer im Wandel befinden (Merger, Verkauf von Unternehmensteilen, Reorganisation, etc.). Deshalb können Archivkonzepte obsolet werden oder die Wartungskosten für ein System in die Höhe schießen.

10.7.4 (№28) Expedia – Prozessoptimierung bringt 25fachen ROI

Anwender	Eddie Satterly, Senior Director of Infrastructure Architecture and Emerging Technologies bei Expedia
Anbieter	Splunk Olav Strand, ostrand@splunk.com
Problem	Akquisitionen und organisches Wachstum haben zu einer erhöhten Komplexität der IT-Umgebung von Expedia, dem weltweit größten Online Reise-Anbieter, geführt. Mit der zunehmenden Unternehmensgröße ist auch die Anzahl der Tools gewachsen, die über die Leistungsfähigkeit der Systeme Aufschluss geben sollten. So musste sich das Team von Expedia plötzlich mit 20 verschiedenen Lösungen auseinandersetzen. Die eingesetzten Lösungen bestanden sowohl aus bewährten Branchenlösungen als auch aus selbst entwickelten Tools. Das Problem: Keine der Lösungen war in der Lage, miteinander zu kommunizieren oder Daten untereinander auszutauschen.
Lösung	Auf der Splunk-Webseite wurde Expedia fündig und holte sich die Splunk-Lösung, um zunächst in einem Pilot-Projekt 1.000 Server zu überwachen. Bereits in der Pilot-Phase konnte die Fehler-Ursachen-Analyse (Root Cause Analysis – RCA) dramatisch verbessert werden, d.h. das Team konnte System- und Prozessfehler binnen kurzer Zeit identifizieren, diagnostizieren und erfolgreich beheben. Die Nachfrage nach Splunk wurde anschließend im gesamten Unternehmen extrem hoch.
Big-Data-Merkmale	Aus dem kleinen Pilot-Projekt wurde ein unternehmensweiter Einsatz: Expedia indiziert und überwacht heute mit Splunk weit mehr als 800 verschiedene Datentypen und über 220.000 Datenquellen für die Fehler-Ursachen-Analyse, Performance-, Web- und Business-Analysen. So wurden innerhalb weniger Monate bereits mehr als 139 Milliarden Ereignisse mit der Splunk-Lösung analysiert. Expedia konnte seine Daten konsolidieren und im Jahre 2011 nahezu 200 Server stilllegen.
Nutzen	Es konnten weitere Vorteile erzielt werden: ■ Kapitalkosten konnten eingespart werden, da kein neues Equipment benötigt wurde. ■ Die Kosten für das Personal, das Management und die Energie des Rechenzentrums konnten drastisch reduziert werden. ■ Softwarelizenzen (Microsoft SQL Server, Betriebssysteme) konnten abgeschafft werden. ■ Durch die Konsolidierungsmaßnahmen konnten Tool-Lizenzen abgeschafft werden. ■ SAN-Speicherkosten konnten reduziert werden.

10.7.5 (№29) NetApp – Diagnoseplattform

Anwender

NetApp, Sunnyvale/Kalifornien, USA
Tel.: +49 (0)89 9005940
www.netapp.com/de

Anbieter

NetApp Deutschland GmbH, Sonnenallee 1,
D-85551 Kirchheim bei München
Tel.: +49 (0)89 9005940
info-de@netapp.com | www.netapp.com/de

Problem

NetApp, Anbieter von Lösungen für innovatives Storage- und Datenmanagement, betreibt eine Diagnoseplattform, mit deren Hilfe Kunden die bei ihnen eingesetzten NetApp-Storage-Systeme auf Performance, Effizienz und Zustand analysieren können. Über die NetApp-Auto-Support-Plattform optimieren Kunden den Einsatz ihrer Storage-Systeme und erhöhen auch die Datensicherheit. Gleichzeitig nutzt NetApp das System zur kontinuierlichen Verbesserung des Kunden-Supports.

Die weltweit installierten NetApp-Storage-Systeme der FAS-Serie senden unstrukturierte Diagnosedaten an das technische Support-Center von NetApp. Die AutoSupport-Plattform empfängt so mehr als 900.000 Datensätze pro Woche mit jeweils rund drei bis fünf Megabyte. Die Datenbank umfasst mehr als ein Petabyte und wächst monatlich um rund sieben Terabyte. Umfangreiche Abfragen über 24 Milliarden Datensätzen hinweg benötigten bis zu vier Wochen, Mustererkennungen über 240 Milliarden Datensätze konnten bislang nicht in einem zeitlich akzeptablen Zeitfenster abgeschlossen werden.

NetApp suchte daher nach einer Lösung, um die riesigen Datenmengen zeitnah zu analysieren, da nur durch eine schnelle Auswertung die Kunden ihre Storage-System bestmöglich verwalten, optimieren und wirtschaftlich nutzen können. Darüber hinaus kann der Kunden-Support nur mit aktuellen und stets verfügbaren Daten zielgerichtet auf aktuelle Anforderungen der Kunden reagieren.

Lösung

Auf der Suche nach einer Lösung evaluierte das NetApp-Team mehr als zehn Big-Data-Systeme. Zu den Kriterien zählten beispielsweise größtmögliche Skalierbarkeit, höchste Performance, umfangreiche Analysefunktionen sowie ein klarer Return on Investment. Schnell wurde klar, dass nur ein System auf Basis von Apache Hadoop in der Lage war, die hohen Anforderungen zu erfüllen. In einer weiteren Analyse wurden die ausgewählten Systeme auf Funktionen wie Parsing, ETL oder Data-Warehouse-Eignung getestet.

Nach Abschluss der Testreihen inklusive einem Proof of Concept entschied sich das Team für die hauseigene Lösung »NetApp Open Solution for Hadoop«. Sie überzeugte durch die größte Geschwindigkeit bei der Auswertung riesiger Datenmengen und bot gleichzeitig die niedrigsten laufenden Betriebskosten (TCO). Implementiert wurden ein Hadoop-Cluster mit 28 Knoten, vier NetApp E2600 Storage-Systeme sowie ein NetApp FAS2040 System.

Big-Data-Merkmale

Immer mehr Unternehmen integrieren Apache Hadoop in ihre Lösungen zur Verarbeitung riesiger verteilter Datenmengen, da nur so eine Echtzeitanalyse großer Datenbestände wirtschaftlich sinnvoll realisierbar ist. Das von NetApp implementierte System trennt zudem Computing und Storage, so dass weniger Rechnerleistung für Storage-bezogene Aufgaben benötigt wird.

Nutzen

Eine Analyse von 24 Milliarden Datensätzen erfolgt heute in nur noch 10,5 Stunden, bislang wurde bis zu vier Wochen dafür benötigt. Eine komplexe Mustererkennung über 240 Milliarden Datensätze führt das neue System in 18 Stunden durch. Somit ist sichergestellt, dass die Gesamtlösung auch bei den weiterhin zu erwartenden Wachstumsraten performante Analysen wirtschaftlich durchführt.

Lessons learnt	Das auf Apache Hadoop basierende System arbeitet sicher, zuverlässig und höchst performant. Die Java-basierende Plattform verwendet offene Technologien und ist somit flexibel erweiterbar. Kunden vermeiden so ein Vendor Lock-in bei gleichzeitig niedrigen Betriebskosten (TCO). Die Storage-Technologien von NetApp ergänzen sich hierzu ideal: Die NetApp E-Serie bietet beispielsweise höchste Skalierbarkeit und wächst mit dem Datenaufkommen, ohne dass die Performance darunter leidet. Die FAS Storage-Systeme speichern die Hadoop NameNodes und schützen HDFS-Metadaten. Insgesamt wird so die Verfügbarkeit der Hadoop-Cluster gesteigert, ein Neuaufbau von Clustern nach einem Plattenwechsel erfolgt sehr rasch und auch die Wiederherstellung nach NameNode-Ausfällen wird deutlich beschleunigt.

10.7.6 (№30) Otto-Gruppe – Mehr Sicherheit und Qualität für die gesamte IT

Anwender	Michael Otremba Abteilungsleiter Softwareentwicklung Otto Gruppe
Anbieter	Splunk Olav Strand, ostrand@splunk.com
Problem	2008 stand der Group Technology Partner (GTP), der IT-Service Provider der Otto Gruppe, vor der Herausforderung, die bestehenden Monitoring Systeme zu harmonisieren und einen konsolidierten Ansatz zur Überwachung und Analyse ihrer aus verschiedenen Datenbanken, Application Servern und Client Applikationen bestehenden Infrastruktur zu initiieren. Mit der Operational Intelligence-Lösung von Splunk fand GTP genau das richtige Tool, um diese Herausforderung zu meistern.
Lösung	Der Erfolg dieses Projekts veranlasste GTP, in einem zweiten Schritt die Splunk-Lösung auch in der Entwicklungsabteilung und der Testumgebung für die Qualitätssicherung einzusetzen, um Code, Konfigurationen und Installationen vor der Produktion zu optimieren. Im dritten Schritt wurde Splunk dann als Monitoring- und Analyse-Lösung für das in-house entwickelte und auf einem zentralen SOA-basierten Server laufendem CRM- und Bestellsystem, die NOA (Neue Otto Anwendung), eingesetzt. NOA erfasst alle Kunden-, Produkt- und Bestellinformationen und ist so eines der geschäftskritischen Systeme des Konzerns. Eine 24/7-Verfügbarkeit von NOA ist daher für die Otto Gruppe zwingend notwendig.
Big-Data-Merkmale	Aus einem kleinen Pilot-Projekt entwickelte sich Splunk bei der Otto Gruppe zu einem zentralen Tool. Heute baut die Otto Gruppe auf Splunk, um seine Infrastruktur, CRM Call Center-Applikation, sein Central Processing System sowie Syslog-Dateien von Switches und internen Proxy-Servern zu überwachen, zu analysieren und zu optimieren. So werden mit Splunk in 24 Stunden im GTP-Umfeld 350 GB an Daten indiziert.
Nutzen	Die Otto Gruppe kann nun ■ die Zeit für die Analyse der Systeme signifikant reduzieren, die Ausfallzeiten verringern und die Instandsetzungszeit verkürzen, ■ viel schneller viel stabilere Lösungen und Installationen einführen und dazu noch die Qualität signifikant verbessern, ■ alle Systeme seiner Call Center in Deutschland in Echtzeit überwachen, einschließlich aller relevanten Backend-Anfragen. So können Fehler und mögliche Ausfallzeiten frühzeitig erkannt, Ursachen dafür identifiziert und Lösungen gefunden werden. ■ Den Zeit- und Ressourcenaufwand für die Pflege und Wartung der Applikation erheblich reduzieren.



10.7.7 (№31) Europäisches Patentamt – Patentrecherche weltweit

Anwender	Europäisches Patentamt
Anbieter	Empolis Information Management GmbH Martina Tomaschowski (martina.tomaschowski@empolis.com)
Problem	Von der Erfindung zum Patent ist es mitunter ein steiniger Weg: Hält der Antragsteller die erforderlichen Richtlinien ein? Gibt es bereits eine identische patentierte Erfindung? Korreliert der Patentantrag auch wirklich mit dem Patentrecht? Zur Erteilung eines Patentes benötigt der Prüfer des Patentamtes Informationen wie bestehende Patente oder Fachliteratur. Diese zu recherchieren kann sich zu einer Herkulesarbeit entwickeln: Informationen sind lokal nicht verfügbar, müssen teilweise von auswärts angefordert werden, und das Herunterladen dauert aufgrund der Serverüberlastung eine Ewigkeit.
Lösung	Die Lösung EPOQUE Net unterstützt die hochperformante Fachrecherche in Patentdatenbanken – und darüber hinaus. EPOQUE Net wurde in enger Zusammenarbeit mit dem Europäischen Patentamt entwickelt und ist heute in 45 Ländern der Erde im Einsatz. Dazu zählen neben den Mitgliedsstaaten der EU auch die Patentämter Brasiliens, Kanadas und Australiens.
Big-Data-Merkmale	**Volume:** Das hohe Speichervolumen von EPOQUE Net erlaubt die problemlose Verarbeitung beliebiger unstrukturierter Texte und einer großen Zahl an Bildern. Mehr als 6.500 Patentprüfer des Europäischen Patentamtes haben Zugang zu den erforderlichen Prüfdaten. Aufgrund der programmierten Zeitvorgaben für einen Prüfprozess kann das System eine beliebige Seite in weniger als 0,3 Sekunden aus dem Archiv anfordern und anzeigen – auch im Hochlastbetrieb. Die dazu eingesetzten Datenbanken sind in ihrem Umfang ständig erweiterbar; derzeit besitzen sie ein Datenvolumen von insgesamt 13 Terabyte mit 450 Millionen Einträgen. Das Online-Archiv enthält gegenwärtig Daten von rund 40 Terabyte. Darüber hinaus lassen sich weitere Datenbanken zwecks Recherche oder Archivierung integrieren. **Variety:** Neben den Patentinformationen werden zunehmend weitere Quellen wie wissenschaftliche Publikationen verwendet, um bereits veröffentliche Technologien zu identifizieren – die so genannte Non-Patent Literature. Die Anlieferung erfolgt in verschiedenen Datenformaten. Insbesondere die Metadaten müssen für die Patentrecherche aufbereitet werden (z. B. Zuordnung zu Patentklassifikationen). Die Patente und Patentanmeldungen können zusätzlich noch in verschiedenen Sprachvarianten vorliegen. **Velocity:** Alle zwei Minuten wird in Europa ein neues Patent angemeldet[72]. In einer Anmeldung sind dabei verschiedene Aspekte und Ansprüche enthalten, welche eigenständig zu recherchieren sind. Pro Arbeitstag werden weiterhin – nach einer Schätzung von 2002 – ca. 20.000 Beiträge im technisch-naturwissenschaftlichen Bereich und damit potenzielle Non-Patent Literature veröffentlicht[73].
Nutzen	Aufgrund seiner Performanz, Schnelligkeit und Benutzerfreundlichkeit beschleunigt EPOQUE Net nachweisbar Recherche und Bearbeitung von Prüfdaten. Effizienz und Leistung der Software haben sich mittlerweile herumgesprochen: Mehr als 70 Prozent der weltweiten Patentanmeldungen werden mit dem Patent-Recherche-System »EPOQUE-Net« geprüft und bearbeitet.
Lessons learnt	Ergonomie und Leistungsfähigkeit waren die Richtlinien, nach denen EPOQUE Net konzipiert wurde: Das System wurde auf der Basis einer Mehrschichtenarchitektur realisiert, um den hohen Anforderungen an Performanz und Skalierbarkeit gerecht zu werden. Unix-Server bilden eine Middleware; ein Cluster von Linux-Servern mit den EPO-eigenen Datenbanken bildet das Backend. Die Datenbanken, von Empolis vollständig als XML-Datenbanken mit UNICODE implementiert, sind binärkompatibel und können auf verschiedenen Plattformen eingesetzt werden. Durch die schnelle Bearbeitung dieser Patentanmeldungen wird ein gesamtwirtschaftlicher Vorteil geschaffen.

72 http://www.epo.org/news-issues/news/2012/20120117_de.html

73 http://www2.fkf.mpg.de/ivs/literaturflut.html

10.7.8 (N°32) Schweizer Staatssekretariat für Wirtschaft – Kostengünstige Höchstleistung für die Schweizer Arbeitsmarktstatistik

Anwender	Staatssekretariat für Wirtschaft SECO, Ressort Arbeitsmarktstatistik, Schweiz. Das SECO ist das Kompetenzzentrum der Schweiz für alle Kernfragen der Wirtschaftspolitik. Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra Swiss Confederation Federal Department of Economic Affairs FDEA **State Secretariat for Economic Affairs SECO**
Anbieter	EMC Computer Systems AG (Greenplum – The Big Data Division of EMC), Informatica für ETL, Microstrategy 9 als BI-Frontend, saracus consulting AG (Beratung).
Problem	Bereits 2007 startete das SECO das Projekt Labor Market Data Analysis (LAMDA X). Im Rahmen von LAMDA X wurden bis 2009 verschiedene BI-Anwendungen mit teils komplexen Berechnungen realisiert, darunter etwa die offizielle Arbeitsmarktstatistik, Auszahlungsstatistiken der Arbeitslosenversicherungen, Führungskennzahlen (KPI) für die RAV-Leiter (RAV=Regionale Arbeitsvermittlung) sowie Arbeitsloseninformationen für die Öffentlichkeit. Doch komplexere Auswertungen riefen 2011 nach einer neuen Infrastruktur, da in der bestehenden Systemarchitektur Performanceprobleme nur mit ziemlichem Aufwand zu bewältigen gewesen wären. Einschränkende Faktoren waren die physische Trennung von Datenbanksystem und Daten sowie der Datentransport über das Netzwerk, das auch von anderen Teilnehmern benutzt wird. Aus diesen Gründen suchte das SECO 2011 nach einer neuen Datenbanklösung für die Arbeitsmarktstatistik, die den Anforderungen von Big Data genügt.
Lösung	Im Dezember 2011 fiel der Entscheid für eine Proof-of-Concept-Installation, in der die EMC Greenplum Database das bisherige Datenbanksystem ersetzt. Die Greenplum-Datenbank arbeitet mit massiv-paralleler Verarbeitung (MPP). Ein Master-Server steuert die Verarbeitung auf beliebig vielen Segment-Servern – die Lösung ist flexibel skalierbar und lässt sich auf unterschiedlichen Hardwareplattformen betreiben. Der Proof of Concept der Arbeitsmarktstatistik umfasst einen Master-Server, ergänzt durch einen zweiten Server für Redundanz, sowie vier Segment-Server.
Big-Data-Merkmale	Besonders die Flexibilität war bei der Entscheidung für EMC Greenplum entscheidend. Das Data Warehouse ist mit rund 500 GB, davon 200 für die Data Marts, relativ klein. Das SECO benötigt also vor allem auch Skalierbarkeit nach unten. Außerdem erlaubt EMC Greenplum den Betrieb auf preisgünstigen Industriestandard-Servern, die bei den üblichen Bundeslieferanten unkompliziert beschafft werden können. Charakteristisch sind zudem die Zielgruppen. Zu den rund 900 Nutzern gehören, neben den zuständigen Bundesstellen, Mitarbeitende der 120 regionalen Arbeitsvermittlungszentren (RAV) in den 26 Kantonen sowie der über 40 Arbeitslosenkassen – für die Arbeitsmarktstatistik ergibt sich eine komplexe und heterogene Nutzergemeinde mit unterschiedlichen Anspruchsgruppen.
Nutzen	Die mit dem Proof-of-Concept angestrebte und auch erreichte Performance-Steigerung erhöht die Flexibilität erheblich. So zeigt sich im Frontend eine deutlich bessere Leistung. Auch bei komplexen Abfragen erscheinen die Ergebnisse rasch. Falls es bei einzelnen Auswertungen doch wieder zu Performance-Problemen kommt, hat das SECO genügend Raum für Optimierungen. Der ETL-Teil, wo die Informatica-Plattform zum Einsatz kommt, biete ebenfalls bereits ansprechende Ladezeiten, müsse hingegen im Detail noch auf die neue Datenbank angepasst werden. Extrem schnell geht nun das Kopieren der Datenbank. Wenn das Data-Warehouse-Team früher für Test- und Entwicklungszwecke oder für anspruchsvolle Spezialauswertungen eine separate Datenbank aufbereiten musste, dauerte dies zwei Wochen. Mit der neuen Lösung ist dafür lediglich ein Kopiervorgang erforderlich, der in zwanzig Minuten erledigt ist.
Lessons learnt	Die neue Lösung ermöglicht nun vor allem auch in die Zukunft gesehen eine Auseinandersetzung mit Fragestellungen, an die man sich mit der ursprünglichen Lösung nicht herangewagt hätte, beispielsweise in Verbindung mit der Statistiksoftware .

10.7.9 (№33) Toll Collect – Qualitätssicherung der automatischen Mauterhebung

Anwender	Toll Collect GmbH, Dr. Bernd Pfitzinger
Anbieter	(Eigenentwicklung auf Basis von Standard-Software)
Problem	90% der Mauteinnahmen werden im deutschen Mautsystem automatisch von über 700.000 On-Board-Units erhoben (OBU). Jeder Nutzer kann sich eine OBU in seinem Fahrzeug installieren lassen, anschließend ist Toll Collect für die korrekte Erhebung der Maut auf allen mautpflichtigen Bundesfernstraßen verantwortlich. Das angestrebte Qualitätsniveau liegt bei 1:1000, d. h. unabhängig vom Ort, von der Tageszeit, von Witterungsbedingungen etc. darf nur jede 1000. Erhebung fehlerhaft sein. Die Herausforderung besteht nun darin, in Echtzeit sich neu entwickelnde Fehlerbilder zu erkennen und zu bereinigen. Besondere Herausforderungen ergeben sich aus folgendem Umstand: Das erreichte Qualitätsniveau bedeutet, dass Fehler nur selten auftreten und u.U. auch nur vorübergehend sichtbar sind. Die Herausforderung liegt darin, potentiell fehlerhafte Geräte vor Eintreten des Fehlers zu erkennen und zu reparieren, ohne zu viele funktionierende Geräte fälschlicherweise auszusortieren.
Lösung	Aufbauend auf der normalen Datenverarbeitung der eingehenden Mautfahrten werden die durchlaufenden Daten in Echtzeit anhand vordefinierter Muster auf bekannte Fehler untersucht. Bei Überschreiten von Schwellwerten oder Trends können korrigierende Maßnahmen eingeleitet werden.
Big-Data-Merkmale	Die Analyse der automatisch erhobenen Maut – entsprechend fast 25 Milliarden gefahrene Kilometer je Jahr – stellt eine große Herausforderung dar, wenn in Echtzeit die Daten auf eine Vielzahl von Mustern untersucht werden sollen. Die Sensitivität der Mustererkennung liegt dabei nur wenig oberhalb des Hintergrundrauschens.
Nutzen	Der Nutzen ist neben der erreichten Qualität der automatischen Mauterhebung in internen Berechnungen auch monetär nachweisbar.
Lessons learnt	Qualitätsansprüche auf dem Niveau von 1:1000 oder besser lassen sich bei Diensten, die allen denkbaren Umwelteinflüssen ausgesetzt sind, nur durch eine kontinuierliche und detaillierte Auswertung aller vorhandenen Daten sicherstellen. Die Herausforderung liegt in der Identifikation und Trennung echter Fehler von vorübergehenden statistischen Ereignissen.

10.7.10 (№34) XING AG – Bewältigung schnell wachsender Datenvolumina

Anwender	XING ist ein soziales Netzwerk für berufliche Kontakte. Über 12 Mio. Mitglieder nutzen die Plattform für Geschäft, Beruf und Karriere. Gerade in der New-Media Branche fallen Unmengen an Daten an. Für interne Analysen verarbeitet XING mehr als zehn Milliarden Datensätze / bzw. einige zehn Terabytes pro Jahr.
Anbieter	Exasol AG, Neumeyerstraße 48, 90411 Nürnberg Carsten Weidmann, Head of Presales Tel: 0911 23 991 0, Email: Carsten.Weidmann@exsasol.com
Problem	Das Business Netzwerk XING wächst schnell – und mit ihm die Datenmenge. Mit Hilfe eines Standard-Datenbanksystems und diversen Daten-Management-Tools, die bei XING bereits seit 2006 im Einsatz waren, konnten diese Datenmengen nur noch unzureichend verarbeitet werden. Teilweise überschritt die Zeit der Datenverarbeitung eines Tages bereits die 24-Stunden-Marke. Das System war weder ausbaufähig noch skalierbar. Es konnte kein Clustering durchgeführt werden. Die BI-Abteilung von XING war deshalb auf der Suche nach einer leistungsstärkeren Datenbank-Management-Lösung.
Lösung	Nach einem erfolgreichen Proof of Concept integrierte XING in nur vier Wochen mit Hilfe des EXASOL-Teams die EXASolution-Datenbank in die bestehende Infrastruktur zur Datenauswertung. Aktuell werden über den Pentaho Data Integrator sämtliche Daten aus dem XING-Live-System, einer MY SQL Datenbank, in die EXASolution-Datenbank geladen. EXASolution analysiert alle eingehenden Daten, die wiederum in einzelnen Berichten über das MicroStrategy-Frontend ausgegeben werden können. Zur Kampagnensteuerung setzt XING darüber hinaus das Kampagnenmanagement-Tool SAS Campaign ein.
Big-Data-Merkmale	In XING verwalten mittlerweile weltweit über 12 Millionen Mitglieder ihre Kontakte und täglich kommen neue dazu. Ein souveräner Umgang mit Massendaten ist für XING unabdingbar: Derzeit werden ca. 10 Milliarden Datensätze anhand einer Vielzahl von Faktoren berechnet und analysiert. Und das muss schnell gehen, denn als Social-Media-Plattform muss XING ihre Angebote in Echtzeit steuern können. Mit dem Wachstum der Plattform hat sich das Datenvolumen innerhalb weniger Jahre vervielfacht und liegt derzeit bei über 30 TB.
Nutzen	XING nutzt die Hochleistungsdatenbank EXASolution, um Geschäftsfelder, Kundenstrukturen oder die Akzeptanz von Produkten und Features schneller und transparenter auszuwerten. In diesem Zusammenhang konnte auch ein professionelles Empfehlungsmarketing aufgebaut werden. Aufgrund der guten Skalierbarkeit kann jederzeit schnell und flexibel auf die stetig wachsende Datenmenge und Nutzeranforderungen reagiert werden.
Lessons learnt	In kürzester Zeit stellten sich positive Effekte bei XING ein, vor allem eine deutliche Verbesserung bei den Analysen. Prozesse können durch die neue Lösung schneller entwickelt und Ad-hoc Anfragen zügiger beantwortet werden. Es sind keine langen Workarounds mehr notwendig, alle BI-Mitarbeiter nutzen das neue System effektiv. Die Komplexität und die Wartung des Systems wurden merklich verringert. Bei der Arbeit mit der neuen Lösung konnte eine steile Lernkurve seitens der Anwender verzeichnet werden, auch wird spürbar produktiver gearbeitet. Gleichzeitig konnte der Administrationsaufwand deutlich verringert werden. Durch eine phasenweise Migration stellten sich schnell erste Erfolge ein. Unternehmen sollten allerdings darauf achten, genügend Zeitpuffer in BI-Projekte einzuplanen. Gern werden interne und externe Gründe unterschätzt, die einen echten Zeitverzug auslösen können.



11 Abkürzungen und Glossar

- ACID
  Atomicity, Consistency, Isolation, Durability

- Ambient Intelligence
  Vernetzung von Sensoren, Funkmodulen und Prozessoren zur Erleichterung im Alltag. Verwandte Begriffe: Ubiquitous Computing, Pervasive Computing.

- Analytics > Business Analytics

- Anonymisierung
  Veränderung personenbezogener Daten mit dem Ziel, den Personenbezug aufzuheben

- Apache Foundation
  ehrenamtlich arbeitende verteilte Gemeinschaft von Entwicklern zur Förderung von Open-Source-Softwareprojekten

- B2B
  Business-to-business

- BDSG
  Bundesdatenschutzgesetz

- BI > Business Intelligence

- BIRT
  Business Intelligence and Reporting Tools - Open-Source-Projekt der > Eclipse Foundation, das Berichtswesen- und Business-Intelligence-Funktionalität für Rich Clients und Web-Applikationen zur Verfügung stellt.

- BITKOM
  Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V.

- Business Analytics
  Gesamtheit der Methoden, Technologien, Applikationen und Best Practices zur Untersuchung wirtschaftlicher Kennziffern mit dem Ziel, Erkenntnisse für die Unternehmensplanung abzuleiten

- Business Intelligence
  Systematische Sammlung und Auswertung von Daten in elektronischer Form mit dem Ziel, Erkentnisse in unternehmerische Entscheidungen einfließen zu lassen

- CAGR
  Compound Annual Growth Rate

- CEP > Complex Event Processing

- Competitive Intelligence
  Sammlung, Analyse und Bereitstellung von Informationen über Produkte, Wettbewerber, Kunden und Marktbedingungen mit dem Ziel der Entscheidungsunterstützung in Unternehmen.

- Complex Event Processing
  CEP-Lösungen extrahieren kontinuierlich entscheidungsrelevantes Wissen aus unternehmenskritischen Daten und bereiten diese bedarfsgerecht auf.

- CRM
  Customer Relationship Management

- Cross Selling
  Verkauf von sich ergänzenden Produkten oder Services

- Crowdsourcing
  Auslagerung von Aufgaben oder Problemen an freiwillige Mitarbeiter im Internet

- Data Governance
  Teilgebiet der Qualitätskontrolle: umfassendes Management (Bewertung, Verwaltung, Nutzung, Verbesserung, Überwachung, Pflege und Schutz) von Unternehmensdaten.

- Data Mart
  Kopie von Teildatenbeständen eines Data-Warehouse, die mit Blick auf einen Organisationsbereich oder eine Applikation hergestellt wird

- Data Mining
  systematische Anwendung von (mathematisch-statistischen) Methoden zur Mustererkennung in einem Datenbestand

- DBMS
  Database Management System

- DMS
  Dokumentenmanagement-System

- DWH
  Data Warehouse

- EBIT
  Earnings before Income and Taxes

- Eclipse Foundation
  gemeinnützige Gesellschaft, die die Eclipse-Open-Source-Gemeinschaft und deren Projekte leitet

- ECM
  Enterprise Content Management

- ERP
  Enterprise Resource Planning

- ESP > Event Stream Processing

- ETL
  Extract – Transform – Load

- Event Stream Processing
  Gesamtheit von Technologien (Visualisierung, Datenbanken, ..., Complex Event Processing) zur Entwicklung von Eventgetriebenen Informationssystemen (event-driven information systems). Beim ESP geht es um die Identifikation wichtiger Ereignisse in einem Ereignisstrom. ESP ist wichtig in den Bereichen Finanzdienstleistungen, Betrugserkennung, Prozessüberwachung sowie bei Location-based Services

- GPS
  Global Positioning System

- Hadoop
  hochverfügbares, leistungsfähiges Dateisystem zur Speicherung und Verarbeitung sehr großer Datenmengen. Dateien werden auf mehrere Datenblöcke verteilt. Zur Steigerung der Zuverlässigkeit und Geschwindigkeit legt Hadoop mehrfach Kopien von einzelnen Datenblöcken an und unterstützt Dateisysteme mit mehreren 100 Millionen Dateien. Es basiert auf einem Algorithmus von Google sowie auf Vorschlägen des Google-Dateisystems und ermöglicht es, Rechenprozesse mit großen Datenmengen auf Computerclustern durchzuführen.

- HANA
  High-Performance Analytic Appliance, auf der > In-Memory-Technologie basierende Datenbanklösung der SAP

- HBase > Hadoop database

- HDFS > Hadoop Distributed File System

- Hive
  Data Warehouse System für > Hadoop



- IaaS
  Infrastructure as a Service
- In-Memory-Technologie
  beruht auf der Verlagerung von Datenspeicherung, -verarbeitung und -analyse in den Arbeitsspeicher eines Computers. Die Zugriffsgeschwindigkeit auf Daten liegt im Arbeitsspeicher mehr als eine Million mal höher als auf andere Datenspeicher. Somit werden Applikationen um Zehnerpotenzen schneller.
- Internet der Dinge
  Bezeichnet die Verknüpfung eindeutig identifizierbarer physischer Objekte (z. B. Maschinen, Geräte) mit einer virtuellen Repräsentation in einer Internet-ähnlichen Struktur
- Location-based Marketing
  Werbung mit ortsbasierten Diensten (Location-based Services)
- M2M
  Machine-to-Machine
- MapReduce
  von Google Inc. eingeführtes Framework für Berechnungen über große Datenmengen auf Computerclustern
- MIA
  Marktplatz für Informationen und Analysen
- Mikrosegmentierung
  Verkleinerung der Marktsegmente bei der Segmentierung von Zielgruppen
- NoSQL
  Neue Datenbanken, die ohne starres Datenbank-Schema arbeiten und beliebige Datenformate speichern können. Ihre Stärke liegt in der Optimierung auf eine verteilte Architektur, die Abfragen auf sehr großen Datenmengen über viele einfache PCs hinweg verteilt.
- OBU
  On-Board Unit
- OLAP
  Online Analytical Processing
- OLTP
  Online Transaction Processing
- Open Innovation
  Öffnung von Innovationsprozessen gegenüber der Außenwelt (z. B. Kunden, Geschäftspartner) mit dem Ziel, (externes) Wissen für Innovationen zu nutzen
- PLM
  Product Lifecycle Management
- POS
  Point of Sale
- PPV
  Pay per view
- Predictive Analytics
  Gesamtheit statistischer Verfahren (Modellierung, Maschinelles Lernen, Data Mining u.a.) zur Analyse wirtschaftlicher Fakten mit dem Ziel, Prognosen zukünftiger Ereignisse zu gewinnen
- Predictive Modeling
  Teilbereich der Datenanalyse; befasst sich mit der Vorhersage von Ereignis-Wahrscheinlichkeiten und Trends
- Product Lifecycle Management
  Konzept zum Management von Produkten oder Services über deren gesamten Produktlebenszyklus
- Pseudonymisierung
  Ähnlich wie bei der Anonymisierung geht es um die Veränderung personenbezogener Daten mit dem Ziel, den Personenbezug »zu verstecken«. Bei der Pseudonymisierung werden Namen durch Codes ersetzt, um die Identifizierung von Personen unmöglich zu machen oder zumindest deutlich zu erschweren
- RAM
  Random Access Memory



- RFID
  Radio-Frequency Identification

- SaaS
  Software as a Service

- SAN
  Storage-Area Network

- Screen Scraping
  Auslesen von Texten aus Computerbildschirmen

- semantische Analyse
  Teilgebiet der Sprachwissenschaft (Linguistik), befasst sich mit Sinn und Bedeutung von Sprache

- Sentiment-Analyse
  Teilgebiet des Text Mining; automatische Auswertung von Texten mit dem Ziel, Stimmungen zu erkennen – also eine Äußerung als positiv oder negativ zu bewerten

- SEO
  Search Engine Optimization

- SMILA
  industriell eingesetztes Open Source Framework

- SMS
  Short Message Service

- SOA
  Service-Oriented Architecture

- SQL
  Structured Query Language

- STB
  Set-Top Box

- TCO
  Total Cost of Ownership

- Text Mining
  Gesamtheit statistischer und linguistischer Verfahren, mit denen in un- oder schwachstrukturierten Texten Bedeutungsstrukturen ermittelt werden. Ziel ist es, wesentliche Informationen der Texte schnell zu erkennen.

- UIMA
  Unstructured Information Management Architecture

12 Sachwortregister

Aadhaar Program 81
Ambient Intelligence 12
Analyse auf sozialen Graphen 61
Analytics 21
Anonymisierung 45
Apache Foundation 25, 27
Arbeitslosenversicherung
  Schweizerische 90
arvato Systems GmbH 56, 101
Asia-Pacific 49
BDSG 43
Betrugserkennung 42, 79
Big Data Analytics 65
Big-Data-Strategie 15
Bildanalytik 21
BIRT 54
BITKOM e.V. 101
Blog 35
Blue Yonder GmbH & Co. KG 58, 62, 102
Bundesdatenschutzgesetz 43
Business Intelligence 15, 23, 24, 62
  klassische 41
CAGR 48
CEP 29
CERN 37
Cloud Computing 19
Cloudera Inc. 76
Competitive Intelligence 30
Complex Event Processing 28, 80
Computerwoche 30
Connected Navigation Device 77
Controlling 41
Cross Selling 35
Crowdsourcing 37
Crowd-Wissen 76
Dailymotion 54
Dashboard 55
Data Governance 16
Data Mart (Cube) 28
Data Mining 21, 28, 44
  Privacy-Preserving 43, 45
Data Scientist 55
Data Warehouse 24
Data-Driven Company 62
Data-Warehousing 15
Daten
  personenbezogene 49
Datenmenge 17
Datenschutz 16
Datenverarbeitung
  linguistische 25, 29
Datenvielfalt 21
Deanonymisierung 45
Deutsche Welle 54
Deutschland 50
DeutschlandCard GmbH 56
Diagnosetechnologie
  wissensbasierte 71
Digital-Factory-Simulation 39
dm 58
Eclipse Foundation 25, 33
eCommerce 49
Embedded Analytics 9, 14, 61
EMC Deutschland GmbH 90, 102
EMC Greenplum 90
EMC Greenplum Database 90
Empolis Information Management GmbH 68, 71, 89, 101
Energietechnik 71
EPOQUE-Net 89
Erdölverarbeitung 38
etracker Dynamic Discovery 59
etracker GmbH 59
Europa 49
Event Stream 25
Eventual Consistency 23
Exadata 56
Exasol AG 67, 72, 92, 102
EXASolution 67, 92
Expedia 85
Experton Group AG 47, 101
Facebook 21, 35, 37
Fahrzeugflotte 77



Fernüberwachung 36
Finanz- und Risiko-Controlling 41
Flottensteuerung 77
Forrester Research GmbH & Co. KG 20, 101
Fraud Detection 43
Fraud Management 79
Fraud-Mining-Technologie 79
Fraud-Szenarien 91
Fraunhofer IAIS 79, 102
Fraunhofer IGD 78, 102
Fujitsu Technology Solutions GmbH 77, 101
fun communications GmbH 82, 101
Gaming 49
Gartner 20
Geonames 32
Geschwindigkeit 21
Gesundheitsvorsorge 38
Gesundheitswesen 36, 38
GPS 69
Graf von Westphalen Rechtsanwälte Partnerschaft 101
Graphen
  soziale 61
Hadoop 27, 28, 29, 54, 76, 81, 83, 86
Hadoop-Cluster 55, 82
HANA 65, 74
HBase 54
HDFS 86
Hewlett-Packard GmbH 66, 101
Hive 82
Hochleistungsdatenbank 92
IaaS 47
IBM Canada 70
IBM Deutschland GmbH 101
IBM Research 69
IBM Schweden 69
Informationssicherheit 16
InfoSphere Streams 70
In-Memory-Technologie 27, 65, 80
In-Store-Ansprache 36
Internet der Dinge 38
Japan 49
Java 86
Kampagnenmanagement 92
Klumpenrisiko 42
Königliches Technologie Institut 69
Kredit
  Ausfallwahrscheinlichkeit 42
Kreditkartenbetrug 79
Kreditkartensicherheit 78
Kreditrisikofaktoren 42
Large Hadron Collider 37
Linguistik 31
Location-based Marketing 36
M2M 11, 38
Machine-Learning Classifyer 54
Machine-to-Machine 12
Macy´s 60
Manipulationsprävention 42
MapReduce 27, 61, 82
Marketing
  Point-of-Sales-basiertes 36
Maschinenbau 71
MasterCard 43
Media-Asset-Management-System 54
Messdaten-Archivierung 72
MIA – Marktplatz für Informationen und Analysen 30
Mikrosegmentierung 35
MINTify rule 79
Mitbewerberanalyse 31
Musteranalyse 61
Mustererkennung 78
MyVideo 54
Mzinga 61
NetApp Deutschland GmbH 86, 102
NeuroBayes® 58
NoSQL 27
OLAP 32
On-Board-Unit 91
Open Innovation 39
Open Linked Data 32
Open Source Framework 33
Operational Intelligence 28
Optimierungsalgorithmus 21
Oracle 56
Otto 62
Otto-Gruppe 88
Parallelisierung 61
ParStream GmbH 59, 64, 102



Patent 89
Patentamt
Europäisches 89
Patent-Recherche-System 89
Pay Per View 63
Paymint AG 79
Pharmaentwicklung 36
Plattform
massiv-parallele 61
PLM 39
PPV 63
Predictive Analytics 28, 58, 62
Predictive Maintenance 39
Predictive Modeling 37
Privacy-Preserving Data Mining 45
Proof of Concept 72, 92
Pseudonymisierung 45
Public Cloud 47
Qualitätssicherung 39
Rapid Prototyping 37
Realtime 21
Remote Service 71
RFID 38
SaaS 47
SAN 85
SAP Deutschland AG & Co. KG 65, 74, 101
SAS Institute 60, 80
SAS Institute GmbH 101
Satelliten-TV 63
Schukat Electronic 65
Schweizer Staatssekretariat für Wirtschaft 90
Scoring 42, 44
Screen Scraping 35
SEA-Analyse 59
Search Analytics 64
Searchmetrics GmbH 64
Semantik 31
semantische Analyse 25
Semikron GmbH 72
Sensorik 21
Sentiment-Analyse 36
SEO 64
SEO Analytics Software 64
Set-Top Box 63
Sevenload 54
Simulation 41
Smartphone 25
SMILA 33
SMS 66
SOA 88
Social Analytics 64
Social Intelligence 61
Social Media 21, 49
Social Networking 12
Software AG 101
Spam- und Phishing-Mails 82
Splunk 82, 85, 88
Splunk Services Germany GmbH 102
SQL 24, 27
Statistical Process Control 72
STB 63
Streams-Computing-Lösung 70
Suchmaschinen-Optimierung 64
System
geschäftskritisches 88
Szenarienbildung 41
TCO 74
Telecom Italia 66
Telematik 71
Telematikdienst 77
Teradata Aster 61
Teradata GmbH 101
Tesco 44
Text Mining 28
Textanalytik 21
The unbelievable Machine Company GmbH 54
THESEUS 32, 33
Time-to-Market 37
Toll Collect GmbH 91, 101
TomTom Business Solutions 77
Total Cost of Ownership 16
Treato 76
Trivadis 56
T-Systems International GmbH 75, 84, 101
Twitter 25, 37
UIMA 33
Unique Identification Authority of India 81
United Overseas Bank 80



University of Ontario 70
USA 49
Vaillant 74
Value-at-Risk-Berechnung 42
Vendor Lock-in 86
Verarbeitung
  massiv-parallele 90
Verfahren
  statistisches 21
Verkehrsmanagement 69
Verkehrstelematik 40
Vertrauen 49
Vorhersagemodell 21
Web Analytics 64
Webtrekk GmbH 67
Web-Video-Plattform 54
Wettbewerbsbeobachtung 35
Wiki 35
Wipro Technologies Germany 63, 101
Wissensgesellschaft 32
Wissensmodell 31, 32
www.mia-marktplatz.de 30
XING AG 92
Yahoo 27
YouTube 54
Zahlungsdienstleister 44
Zeitreihenanalyse 61
Zielgruppenmarketing 36



Autoren des Leitfadens

Jörg Bartel, Senior Software Client IT Architect, IBM Deutschland GmbH

Arnd Böken, Partner, Graf von Westphalen Rechtsanwälte Partnerschaft

Björn Decker, Produkt Manager IAS, Empolis Information Management GmbH

Guido Falkenberg, Vice President Enterprise Transaction Systems, Software AG

Robert Guzek, ETERNUS Business Management Germany, Fujitsu Technology Solutions GmbH (Redaktionsteam)

Steve Janata, Senior Advisor & Channel Program Manager, Experton Group AG

Dr. Thomas, Keil, Program Marketing Manager Business Analytics, SAS Institute GmbH (Redaktionsteam)

Dr. Holger Kisker, Principal Analyst, Forrester Research GmbH & Co. KG

Ralf Konrad, Offering Manager Enterprise Information Management, T-Systems International GmbH

Wulf Maier, Leiter der Enterprise Information Solution Practice (DACH&CEE), Hewlett-Packard GmbH

Axel Mester, Client Technical Architect, IBM Deutschland GmbH

Boris Andreas Michaelis, Head of Consulting Area Business Analytics, SAP Deutschland AG & Co. KG

Mario Palmer-Huke, Regional Practice Manager Analytics & Information Management, Wipro Technologies Germany

Dr. Bernd Pfitzinger, Senior Experte Mautprozesse, Toll Collect GmbH (Redaktionsteam)

Martina Tomaschowski, Vice President Marketing and Public Relations, Empolis Information Management GmbH

Ralph Traphöner, Leiter Technologies, Empolis Information Management GmbH

Jürgen Urbanski, Vice President of Cloud Architectures, Cloud Technologies and Enabling Platforms T-Systems International GmbH

Dr. Carlo Velten, Senior Advisor, Experton Group AG

Dr. Mathias Weber, Bereichsleiter IT-Services, BITKOM e.V. (Redaktionsteam)

Melih Yener, Director of Enterprise Architecture, T-Systems International GmbH (Redaktionsteam)

Die Autoren haben Informationen und Abbildungen ihrer Unternehmen genutzt.
An den Big-Data-Einsatzbeispielen haben mitgewirkt:

Klaas Wilhelm Bollhoefer, Data Scientist, The unbelievable Machine Company GmbH

Johannes Feulner, Geschäftsführer, fun communications GmbH

Norbert Franke, Director Business Development, arvato Systems GmbH

Dr. Jörn Kohlhammer, Abteilungsleiter, Fraunhofer IGD Institut für Graphische Datenverarbeitung

Dr. Michael May, Abteilungsleiter Knowledge Discovery, Fraunhofer IAIS Institut für Intelligente Analyse- und Informationssysteme

Dr. Andreas Ribbrock, Senior Architect, Teradata GmbH

Dunja Riehemann, Director Marketing, Blue Yonder GmbH & Co. KG

Jörg Ruwe, Leiter Sales & Marketing, ParStream GmbH

Florian Seidl, Sales Business Development Manager, NetApp Deutschland GmbH

Olav Strand, Director Germany, Switzerland & Austria, Splunk Services Germany GmbH

Andreas vom Bruch, PR Manager, EMC Deutschland GmbH

Carsten Weidmann, Head of Presales, Exasol AG


BITKOM

Der Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. vertritt mehr als 1.700 Unternehmen, davon über 1.200 Direktmitglieder mit etwa 135 Milliarden Euro Umsatz und 700.000 Beschäftigten. Hierzu gehören fast alle Global Player sowie 800 leistungsstarke Mittelständler und zahlreiche gründergeführte, kreative Unternehmen. Mitglieder sind Anbieter von Software und IT-Services, Telekommunikations- und Internetdiensten, Hersteller von Hardware und Consumer Electronics sowie Unternehmen der digitalen Medien und der Netzwirtschaft. Der BITKOM setzt sich insbesondere für eine Modernisierung des Bildungssystems, eine innovative Wirtschaftspolitik und eine zukunftsorientierte Netzpolitik ein.



Bundesverband Informationswirtschaft,
Telekommunikation und neue Medien e.V.

Albrechtstraße 10 A
10117 Berlin-Mitte
Tel.: 030.27576-0
Fax: 030.27576-400
bitkom@bitkom.org
www.bitkom.org